ANIL SETH

La creación del YO

UNA NUEVA CIENCIA DE LA CONCIENCIA

TRADUCCIÓN DE **ALBINO SANTOS MOSQUERA**

ensayo sexto piso

La creación del yo

La creación del yo

Una nueva ciencia de la conciencia

ANIL SETH

TRADUCCIÓN DE ALBINO SANTOS MOSQUERA

Todos los derechos reservados.
Ninguna parte de esta publicación puede ser reproducida,
transmitida o almacenada de manera alguna sin el permiso previo del editor.

Título original
Being You. A New Science of Consciousness

Copyright © Anil Seth, 2021

Primera edición: 2023

Traducción
Albino Santos Mosquera

Copyright © Editorial Sexto Piso, S. A. de C. V., 2023
América 109
Colonia Parque San Andrés, Coyoacán
04040, Ciudad de México

Sexto Piso España, S. L.
C/ Los Madrazo, 24, semisótano izquierda
28014, Madrid, España

www.sextopiso.com

Diseño
Estudio Joaquín Gallego

Formación
Grafime

ISBN: 978-84-19261-40-3

Para mi madre, Ann Seth, y en memoria
de mi padre, Bhola Nath Seth

El cerebro – es más amplio que el cielo –
pues – colócalos juntos –
y el uno contendrá al otro
con holgura – y a ti – también –

EMILY DICKINSON

PREFACIO

Hace cinco años, por tercera vez en mi vida, dejé de existir. Me estaban realizando una pequeña intervención quirúrgica y mi cerebro empezó a llenarse de anestesia. Recuerdo las sensaciones de oscuridad, de desmoronamiento, de dejarme ir…

La anestesia general es algo muy distinto de dormirse. Tiene que serlo. Si solo estuvieras dormido, el bisturí del cirujano te despertaría al momento. Los estados de anestesia profunda tienen más elementos en común con condiciones catastróficas como el coma o el estado vegetativo, en las que la conciencia está totalmente ausente. Bajo los efectos de una anestesia profunda, la actividad eléctrica del cerebro se paraliza casi por completo, lo cual casi nunca sucede en la vida normal, ni en la despierta ni en la dormida. Uno de los milagros de la medicina moderna es precisamente ese: que los anestesiólogos sepan cómo modificar rutinariamente los cerebros de las personas para que puedan entrar en estados tan inconscientes como esos y regresar de ellos. Es un acto de transformación, una especie de magia: la anestesia es el arte de convertir personas en objetos.

Esos objetos, claro está, vuelven luego a la condición de personas. Así que yo también volví, somnoliento y desorientado, pero ciertamente *en mí*. Parecía que no hubiera transcurrido tiempo alguno. Cuando me despierto de un sueño profundo, a veces me siento confundido respecto a qué hora es, pero tengo siempre la impresión de que, como mínimo, ha pasado un rato, de que ha habido cierta continuidad entre mi conciencia anterior y mi conciencia actual. Bajo los efectos de una anestesia general, la cosa cambia. Yo podría haber estado anestesiado cinco minutos, cinco horas, cinco años…, incluso cincuenta. Y tampoco

diría exactamente que estaba «bajo» sus efectos. Simplemente, yo no estaba allí. Fue como una premonición (extrañamente reconfortante, con aquella ausencia de todo) del olvido total que es la muerte.

La anestesia general no solo actúa sobre tu cerebro o sobre tu mente. Actúa sobre tu conciencia. Al modificar el delicado equilibrio electroquímico interno de los circuitos neuronales en tu cabeza, el estado fundamental básico de tu sentido de «ser» algo o alguien queda (temporalmente) abolido. En ese proceso radica uno de los mayores misterios que la ciencia –y también la filosofía– tiene aún por resolver.

Hay algo que hace que, dentro de cada uno de nuestros cerebros, la actividad combinada de los miles de millones de diminutas máquinas biológicas que son las neuronas dé origen a una experiencia consciente. Y no hablamos de una experiencia consciente cualquiera, sino de *tu* experiencia consciente, aquí y ahora. *¿Cómo sucede algo así? ¿Por qué experimentamos la vida en primera persona?*

Tengo el recuerdo de infancia de haberme mirado en el espejo del baño y haberme dado cuenta por vez primera de que mi experiencia en aquel preciso momento –la experiencia de *ser yo*– terminaría en algún momento, y de que «yo» me moriría. Debía de tener ocho o nueve años por entonces y no es un recuerdo muy fiable (ningún recuerdo temprano lo es). Pero quizá fue en ese momento cuando me percaté también de que, si mi conciencia podía llegar a su fin, tenía que estar ligada de algún modo a la materia de la que estaba hecha, a la materialidad física de mi cuerpo y de mi cerebro. Tengo la sensación de que, desde entonces, he estado tratando de resolver ese misterio de un modo u otro.

Cuando era un estudiante de grado en la Universidad de Cambridge a principios de los años noventa, sumé a mi amor adolescente por la física y la filosofía la fascinación por la psicología y la neurociencia, pese a que, por aquel entonces, ambos campos parecían evitar –cuando no proscribir, incluso– toda alusión a la conciencia. La investigación para mi tesis doctoral me hizo tomar un largo e inesperadamente valioso desvío por la inteligencia artificial y la robótica, seguido de una estancia

de seis años en el Instituto de Neurociencias de San Diego, a orillas del Pacífico, que fue lo que por fin me brindó la oportunidad de investigar directamente la base cerebral de la conciencia. Allí trabajé con el premio nobel Gerald Edelman, una de las figuras que más significativamente han contribuido a que la conciencia vuelva a considerarse un foco de atención legítima para la ciencia.

En la actualidad, hace más de una década que soy codirector de un centro de investigación –el Centro Sackler de la Ciencia de la Conciencia, de la Universidad de Sussex– emplazado entre las suaves y verdes colinas de los South Downs, en el extremo sur de Inglaterra, junto a la ciudad costera de Brighton. Nuestro centro reúne a neurocientíficos, psicólogos, psiquiatras, expertos en neuroimagen, magos de la realidad virtual, matemáticos y filósofos; todos unidos por el objetivo común de abrir nuevas puertas de acceso a la base cerebral de la experiencia consciente.

La conciencia es un misterio que importa tanto a los que somos científicos como a quienes no. Para cada uno de nosotros, nuestra experiencia consciente es *lo único que hay*. Sin ella, no hay nada en absoluto: ni mundo, ni yo, ni interior, ni exterior.

Imagina que una versión futura de mí mismo, tal vez no muy lejana en el tiempo, te ofrece el trato de tu vida: puedo sustituirte el cerebro por una máquina que sea igual a él en todos los sentidos, de tal forma que, desde fuera, nadie pueda notar la diferencia. Esta nueva máquina presenta múltiples ventajas: es inmune al deterioro y puede que incluso te permita vivir para siempre.

Pero hay una salvedad. Como ni siquiera ese yo mío del futuro está seguro de cómo se origina la conciencia a partir de los cerebros reales, no te puedo garantizar que vayas a tener experiencias conscientes si aceptas esta oferta. Tal vez las tengas si la conciencia depende únicamente de la capacidad funcional, del poder y la complejidad de los circuitos del cerebro, pero quizá no las tengas si la conciencia depende en realidad de un material biológico específico (las neuronas, por

ejemplo). Por supuesto, como tu cerebro-máquina produce comportamientos idénticos a los anteriores en todos los sentidos, cuando le pregunto a tu nuevo tú si estás consciente, tu nuevo tú dirá que sí. Pero ¿y si, a pesar de esa respuesta tuya, la vida –para *ti*– ya no es una experiencia en primera persona?

Sospecho que no aceptarías el trato. Sin conciencia, puede que difícilmente importe si vives cinco años más o quinientos. En todo ese tiempo, *no existiría un modo de ser tú.*

Juegos filosóficos aparte, es fácil entender la importancia práctica de conocer la base cerebral de la conciencia. No se puede negar que la anestesia general es uno de los más grandes inventos de todos los tiempos. Mucho menos agradables son las angustiantes perturbaciones de la conciencia que pueden acompañar a ciertas lesiones cerebrales y enfermedades mentales para el creciente número de personas –entre las que me incluyo– que viven de cerca esos problemas. Además, para todos y cada uno de nosotros, las experiencias conscientes cambian a lo largo de toda la vida: pasan de la confusión floreciente y frenética de nuestros primeros años a la clarividencia (probablemente ilusoria y, desde luego, no universal) de la adultez, para desembocar en una deriva final hacia la disolución progresiva –y, para algunas personas, desorientadoramente rápida– del yo por causa del declive neurodegenerativo de la vejez. En cada fase de ese proceso, tú existes, pero la noción de que hay un único yo consciente (¿un alma?) que persiste a lo largo del tiempo podría no ser más que un burdo error. De hecho, uno de los aspectos más llamativos del misterio de la conciencia es la naturaleza del *yo.* ¿Se puede tener conciencia sin autoconciencia, y, en ese caso, seguiría importando igual?

Las respuestas a preguntas difíciles como esas tienen múltiples implicaciones en lo que respecta a cómo concebimos el mundo y la vida que este contiene. ¿Cuándo empieza a haber conciencia en el desarrollo de una persona? ¿Surge ya en el momento de nacer o incluso antes, cuando aún está en el seno materno? ¿Y qué podemos decir de la

conciencia en los animales no humanos, y no ya en los primates y en otros mamíferos, sino también en criaturas tan «de otro mundo» como el pulpo, y quién sabe si hasta en organismos simples como los nematodos o las bacterias? ¿Hay un «modo de ser» una *Escherichia coli* o una lubina? ¿Y las máquinas del futuro? En este sentido, debería interesarnos no solo el poder que las nuevas formas de inteligencia artificial están adquiriendo sobre nosotros, sino también si (y cuándo) tendremos *nosotros* que adoptar una actitud ética hacia *ellas.* En mi caso, estas preguntas evocan la extraña compasión que sentí cuando vi a Dave Bowman destruir la personalidad de HAL en la película *2001. Una odisea del espacio* mediante el simple acto de extraerle sus bancos de memoria, uno por uno. Y en la empatía, más extendida aún, despertada por el sufrimiento de los replicantes de Ridley Scott en *Blade Runner* hay también un buen indicio de la importancia que nuestra naturaleza como *máquinas vivas* tiene para la experiencia de ser un yo consciente.

Este es un libro sobre la neurociencia de la conciencia, que es un intento de entender cómo se relaciona el universo interior de la experiencia subjetiva con los procesos biológicos y físicos que se despliegan en los cerebros y los cuerpos, y cómo puede explicarse aquella en función de estos. Este es el proyecto que ha acaparado mi interés durante toda mi vida profesional, y creo que actualmente hemos alcanzado un punto en el que empiezan a vislumbrarse destellos de posibles respuestas.

Son destellos que ya están cambiando (radicalmente, además) nuestro modo de pensar acerca de las experiencias conscientes del mundo que nos rodea, y de nosotros mismos como parte de él. La forma en que concebimos la conciencia afecta a todos los aspectos de nuestras vidas. Una ciencia de la conciencia equivale nada menos que a una explicación de quiénes somos, de qué significa «ser yo» (o «ser tú») y de por qué se «es» algo, para empezar.

La historia que contaré aquí es una visión personal, modelada por muchos años de investigación, contemplación y conversación. Tal como yo la entiendo, la conciencia no es un enigma que se vaya a «resolver»

igual que se descodificó el genoma, o se determinó que el cambio climático es real. Tampoco sus misterios se revelarán de pronto ante un golpe de inspiración repentino, un momento «eureka» (que, por cierto, no deja de ser un mito sobre la progresión del conocimiento científico tan simpático como, a menudo, alejado de la realidad).

Para mí, una ciencia de la conciencia debería explicar de qué modo las diversas propiedades de la conciencia dependen del funcionamiento de la biología neuronal interna (el *wetware,* por contraposición a *hardware* y a *software)* de nuestra cabeza, y se relacionan con él. Una ciencia de la conciencia no debería tener por objetivo (o, cuando menos, no por objetivo principal) explicar por qué la conciencia forma parte de nuestro universo como un fenómeno más del mismo ni tampoco conocer el funcionamiento del cerebro en toda su complejidad al tiempo que barremos bajo la alfombra el misterio de la conciencia. Lo que espero mostrarte es que, explicando las propiedades de la conciencia en términos de mecanismos cerebrales y corporales, los porqués y los cómos metafísicos profundos de la conciencia se vuelven lenta y progresivamente menos misteriosos.

Empleo la palabra *wetware* (traducible como «soporte húmedo») para subrayar que los cerebros no son simples ordenadores hechos de carne. Son máquinas químicas en la misma medida en que son redes eléctricas. Todos los cerebros que han existido han formado parte de un cuerpo vivo, integrado en (y en interacción con) su entorno: un entorno que, en muchos casos, contiene otros cerebros corporeizados. Explicar las propiedades de la conciencia en términos de mecanismos biofísicos obliga a concebir los cerebros –y las mentes conscientes– como sistemas *corporeizados* e *integrados* (en entornos).

En última instancia, quiero dejarte una nueva concepción del yo, ese aspecto de la conciencia que es probablemente el más significativo para cada uno de nosotros. Una influyente tradición, cuyos orígenes se remontan, como mínimo, a Descartes en el siglo XVII, sostiene que los animales no humanos carecen de yoidad consciente porque no poseen

unas mentes racionales que guíen su conducta. Son meros «animales-máquina», autómatas de carne y hueso desprovistos de la capacidad para reflexionar sobre su propia existencia.

Yo no estoy de acuerdo. A mi juicio, la conciencia tiene más que ver con el hecho de estar vivo que con el de ser inteligente. Somos yos conscientes precisamente *porque* somos animales-máquina. Aquí argumentaré que las experiencias de *ser tú,* o de *ser yo,* surgen de la forma en que el cerebro predice y controla el estado interno del cuerpo. La esencia de la yoidad individual no es la mente racional ni el alma inmaterial. Es un *proceso* biológico de una profunda naturaleza corporal, un proceso sobre el que se sustenta la propia sensación elemental de estar vivo, que es la base de todas nuestras experiencias del yo y, de hecho, de todas las experiencias conscientes en general. *Ser tú* es algo literalmente relacionado con tu cuerpo.

Este libro se divide en cuatro partes. En la primera, explico mi manera de enfocar el estudio científico de la conciencia. En esa parte, también abordo la cuestión del «nivel» consciente –de cuán consciente puede ser o estar alguien o algo– y de los progresos realizados en los intentos de «medir» la conciencia. La segunda parte trata el tema del «contenido» consciente, es decir, de qué es aquello de lo que somos conscientes cuando somos conscientes. En la tercera parte, el foco se vuelve hacia el interior, hacia el yo, y hacia las diversas experiencias que implica la yoidad consciente. En la cuarta (y última) parte –dedicada a «los otros»–, se explora lo que este nuevo modo de entender la conciencia puede decirnos acerca de los demás animales y de la posibilidad de que haya máquinas sintientes. Al acabar el libro, comprenderás que nuestras experiencias conscientes del mundo y del yo son formas de predicción con una base cerebral –«alucinaciones controladas»– que nacen de, se producen a través de, y *son debidas a,* nuestros cuerpos vivos.

Pese a lo empañada que está su reputación entre los neurocientíficos, no se puede negar que Sigmund Freud tuvo razón en muchas cosas. Freud echó la vista atrás, a la historia de la ciencia, e identificó tres grandes

«golpes» asestados al ego de la especie humana por tres importantes avances científicos que chocaron con una fuerte resistencia en sus momentos históricos respectivos. El primero fue el de Copérnico, que evidenció con su teoría heliocéntrica que la tierra orbita en torno al sol, y no al revés. Nos dimos cuenta así de que no somos el centro del universo, sino una simple mota en medio de la inmensidad, un puntito azul claro suspendido en el abismo. Luego vino Darwin, que reveló que compartimos una ascendencia común con todos los demás seres vivos, una revelación que –por asombroso que parezca– aún hoy en día choca con una fuerte resistencia en ciertas partes del mundo. El tercer golpe al excepcionalismo humano lo dio el propio Freud, proclamaba él sin falsas modestias, con su teoría de la mente inconsciente, que cuestionó la idea de que nuestras vidas mentales están bajo nuestro propio control consciente y racional. Aunque tal vez errara mucho en los detalles, Freud acertó de pleno cuando señaló que, el día en que hubiera una explicación naturalista de la mente y la conciencia, llegaría un nuevo (y tal vez definitivo) empujón que contribuyera al destronamiento de la humanidad.

Estos cambios en el modo de vernos a nosotros mismos son siempre positivos. Con cada nuevo avance en nuestro conocimiento, llega una nueva sensación de asombro y una nueva posibilidad para vernos menos *apartados* del resto de la naturaleza y más como *una parte de* esta.

Nuestras experiencias conscientes forman parte de la naturaleza tanto como nuestros cuerpos y tanto como nuestro mundo. Y cuando la vida termine, también lo hará la conciencia. Cuando pienso en esto, me siento transportado de vuelta a mi experiencia (o, mejor dicho, *«no* experiencia») de la anestesia, a aquel abandono en los brazos del olvido y la inconsciencia: un olvido reconfortante, tal vez, pero olvido, al fin y al cabo. Meditando sobre la mortalidad, el novelista Julian Barnes lo expresó a la perfección: cuando llega el final de la conciencia, no hay nada –absolutamente *nada*– de lo que asustarse.[1]

I. NIVEL

1. EL PROBLEMA REAL

¿Qué es la conciencia?

Para una criatura consciente, *ser* esa criatura es serlo de algún modo. Hay un modo de ser yo, un modo de ser tú y probablemente un modo de ser una oveja o un delfín. Cada una de esas criaturas está teniendo unas experiencias subjetivas. Ser yo es *sentirse de cierto modo.* Pero es casi seguro que no existe un modo de ser una bacteria, una hoja de hierba o un robot de juguete. Para ninguna de esas cosas hay nunca (presumiblemente) ninguna experiencia subjetiva en juego: ningún universo interior, ni estado consciente, ni conciencia.

El filósofo más directamente asociado con esta forma de plantear las cosas es Thomas Nagel, quien en 1974 publicó un artículo ya legendario titulado «¿Cómo es ser un murciélago?», en el que defendía que, aunque los humanos jamás pudiéramos vivir las experiencias de uno de esos roedores voladores, para un murciélago siempre habría un modo de ser un murciélago.[*] Siempre he sentido predilección por el enfoque de Nagel, porque pone el acento en la *fenomenología,* es decir, en las propiedades subjetivas de la experiencia consciente, como, por ejemplo, por qué una experiencia visual tiene la forma, la estructura y las cualidades que tiene en comparación con las propiedades subjetivas de una experiencia emocional o con las de una experiencia olfativa. En filosofía, a estas propiedades se las llama a veces también *qualia:* la rojez del rojo, la quemazón de los celos, o las punzadas agudas o el suplicio sordo de un dolor de muelas.

Para que un organismo sea consciente tiene que poseer algún tipo de

fenomenología propia. *Toda* clase de experiencia –*toda* propiedad fenomenológica– vale tanto como cualquier otra. Allí donde hay experiencias, hay fenomenología, y donde hay fenomenología, hay experiencias. Una criatura que solo exista un momento será consciente mientras haya un modo de ser ella, aun si lo único que sucede entonces es una efímera sensación de dolor o de placer.

Resulta útil distinguir las propiedades fenomenológicas de la conciencia de sus propiedades *funcionales* y *conductuales*. Estas últimas hacen referencia a los papeles que la conciencia puede jugar tanto en el funcionamiento de nuestras mentes y cerebros como en los comportamientos de los que un organismo es capaz por el hecho de tener experiencias conscientes. Pero, aunque las funciones y las conductas asociadas a la conciencia son temas importantes, no constituyen el mejor terreno en el que buscar definiciones. La conciencia, ante todo, es una experiencia subjetiva: es fenomenología.

Esto puede parecernos obvio, pero no siempre lo fue. En varias épocas pasadas, se ha confundido ser consciente con tener habla, con ser inteligente, o con manifestar conductas de un determinado tipo. Pero la conciencia no depende del comportamiento externo, como bien se pone de manifiesto cuando soñamos o con casos como los de las personas que sufren estados de parálisis corporal absoluta. Además, decir que el habla es una condición necesaria de la conciencia es como afirmar que los bebés, los adultos que han perdido sus facultades lingüísticas y la mayoría (si no la totalidad) de los animales no humanos carecen de conciencia. Y el pensamiento abstracto complejo solo es una pequeña parte –por muy característicamente humano que pueda ser– de la condición consciente en general.

Hay teorías muy destacadas en la ciencia de la conciencia que continúan poniendo el énfasis en la función y la conducta antes que en la fenomenología. Especial relevancia entre ellas tiene la teoría del «espacio de trabajo global», desarrollada a lo largo de muchos años por el psicólogo Bernard Baars y el neurocientífico Stanislas Dehaene, entre

otros.[1] Según esta teoría, el contenido mental (las percepciones, los pensamientos, las emociones, etcétera) se vuelve consciente cuando logra acceder a un «espacio de trabajo» que –en un sentido anatómico– se reparte entre las regiones frontal y parietal del córtex. (El córtex, o corteza cerebral, es la superficie más exterior y replegada del cerebro, formada por neuronas muy apretujadas).[*] Cuando se emite contenido mental dentro de ese espacio de trabajo cortical, es cuando somos conscientes de él y puede usarse para guiar la conducta de formas mucho más flexibles que con la percepción inconsciente. Por ejemplo, yo me percato conscientemente de que hay un vaso de agua sobre la mesa que tengo delante. Podría asirlo y beberme su contenido, o tirar este sobre mi ordenador (tentador), o escribir un poema al respecto, o llevármelo de vuelta a la cocina porque me he dado cuenta de que está ahí desde hace días. La percepción inconsciente, por el contrario, no permite semejante grado de flexibilidad conductual.[2]

Otra teoría destacada de este tipo es la del «pensamiento de orden superior», una de cuyas hipótesis es que el contenido mental se vuelve consciente cuando existe un proceso cognitivo «de orden superior» que está en cierto modo orientado hacia aquel y, de ese modo, lo trae a la conciencia. Según esta teoría, la conciencia está estrechamente ligada a procesos como la *metacognición* –que es «la cognición acerca de la cognición»–, lo que, de nuevo, pone más el énfasis en las propiedades funcionales que en la fenomenología, aunque no tanto como lo hace la teoría del espacio de trabajo global. Como esta, las teorías del pensamiento de orden superior también resaltan el papel de las regiones frontales del cerebro como un aspecto clave para la conciencia.[3]

Aunque se trata de teorías interesantes e influyentes, en este libro no tengo mucho que añadir a propósito de ninguna de las dos, dado que ambas subrayan los aspectos funcional y conductual de la conciencia, mientras que el enfoque que yo adopto parte de la fenomenología –de lo experiencial en sí– y solo desde esta, como base, se formulan luego postulados acerca de las funciones y las conductas.

Admito que definir la conciencia como «cualquier clase de experiencia subjetiva» puede resultar simple e incluso puede parecer trivial, pero eso es bueno. Cuando solo se tiene un conocimiento incompleto de un fenómeno complejo, aventurar prematuramente definiciones precisas puede resultar restrictivo e incluso puede inducir a engaño. La historia de la ciencia ha demostrado en muchas ocasiones que las definiciones útiles son aquellas que evolucionan a la par que la comprensión científica de los fenómenos y que actúan como andamios del progreso científico más que como puntos de partida o como fines en sí mismas. En genética, por ejemplo, la definición de «gen» ha cambiado considerablemente a medida que ha avanzado la biología molecular.[4] De igual modo, a medida que se desarrolle nuestro conocimiento de la conciencia, también evolucionará la definición (o las definiciones) de esta. Si, por el momento, aceptamos que la conciencia tiene que ver ante todo con la fenomenología, ya podemos pasar a la siguiente cuestión.

¿Cómo tiene lugar la conciencia? ¿Cómo se relacionan las experiencias conscientes con la maquinaria biofísica interna de nuestros cerebros y nuestros cuerpos? ¿Cómo se relacionan, en realidad, con el torbellino de átomos o cuarks o supercuerdas o cualquiera que sea el elemento esencial del que, en última instancia, se compone la totalidad de nuestro universo?

La formulación clásica de esta pregunta es lo que se conoce como el «problema difícil» de la conciencia. La expresión en sí fue acuñada por el filósofo australiano David Chalmers a principios de los noventa y ha marcado las prioridades de buena parte de la ciencia de la conciencia desde entonces. Así describió él el mencionado problema:

> Es innegable que algunos organismos son sujetos de experiencias. Pero la pregunta de cómo es que estos sistemas son sujetos de experiencias es complicada. ¿Por qué cuando nuestros sistemas cognitivos proceden a procesar información visual y auditiva, tenemos una experiencia visual o auditiva: apreciamos la cualidad del azul oscuro, la sensación de un do central? ¿Cómo podemos explicar por qué hay algo como tener una imagen

> mental, o como sentir una emoción? Hay un consenso bastante general en torno a la idea de que las experiencias nacen de una base física, pero carecemos de una buena explicación de por qué y de cómo. ¿Por qué el procesamiento físico da origen a una rica vida interior? No parece objetivamente razonable que lo haga, y sin embargo, así es.[5]

Chalmers contrasta este problema difícil de la conciencia con el denominado problema fácil (aunque sería mejor plantearlos en plural: problema*s* fácil*es),* que tiene que ver con la explicación de cómo unos sistemas físicos –como los cerebros– pueden originar una serie cualquiera de propiedades funcionales y conductuales.[6] Entre estas propiedades funcionales se incluyen cosas como el procesamiento de señales sensoriales, la selección de acciones y el control del comportamiento, la atención, la generación de lenguaje, etcétera. Los problemas fáciles abarcan todas aquellas cosas que seres como nosotros pueden hacer y que se pueden especificar en términos de una función –un modo de transformar un *input* en un *output*– o en términos de un comportamiento.

Obviamente, los problemas fáciles no son, ni mucho menos, tan sencillos como su nombre parece indicar. Los neurocientíficos tardarán décadas o siglos en resolverlos. Lo que Chalmers quiso decir es que los problemas fáciles son de sencilla resolución *en principio,* algo que no puede decirse del problema difícil. Concretamente, Chalmers entendió que no existe ningún obstáculo conceptual a que a los problemas fáciles se les termine hallando explicación en términos de mecanismos físicos. No parece así, sin embargo, en el caso del problema difícil, para el que se diría que ninguna explicación podría estar jamás a la altura de resolverlo. (Digamos, a efectos aclaratorios, que un «mecanismo» se puede definir como un sistema de partes causalmente interactivas que producen efectos).[7] Así, aun después de que los problemas fáciles hayan ido cayendo uno tras otro, el problema difícil seguirá ahí, incólume. «Incluso cuando ya hayamos explicado el desempeño de todas las funciones más o menos relacionadas con la experiencia –la

discriminación perceptiva, la categorización, el acceso interno, el relato verbal–, es muy posible que continúe quedando una pregunta más sin responder: *¿Por qué el desempeño de estas funciones está acompañado por la experiencia?».*[8]

Las raíces del problema difícil se remontan a la Antigua Grecia, o puede que incluso a algún momento anterior, pero resultan especialmente visibles en la división del universo en sustancia pensante, o *res cogitans,* y sustancia material, o *res extensa,* que formuló René Descartes en el siglo XVII. Esa distinción inauguró la filosofía del dualismo e hizo que todos los debates y análisis sobre la conciencia fuesen complejos y confusos desde entonces. Y nada patentiza más esa confusión que la proliferación de diferentes marcos filosóficos de reflexión sobre la conciencia.

Respira hondo, que aquí vienen los «ismos».

Mi postura filosófica preferida –que es también el supuesto por defecto del que parten muchos neurocientíficos– es el *fisicalismo.* Es la idea según la cual el universo está hecho de materia física, y los estados conscientes no son más que –o emergen de– disposiciones particulares de esa sustancia física. Algunos filósofos emplean el término *materialismo* en vez de fisicalismo, pero, a los efectos que aquí nos ocupan, ambos pueden usarse como sinónimos.[9]

En el extremo opuesto al fisicalismo, encontramos el *idealismo.* Es la idea –a menudo asociada al obispo George Berkeley, pensador del siglo XVIII– según la cual la fuente última de la realidad es la conciencia (o mente) y no la sustancia física (o materia). El problema en el que se centra no es tanto el de cómo surge la mente de la materia, como el de cómo surge la materia de la mente.

En una precaria posición intermedia encontramos a los *dualistas* como Descartes, que creen que la conciencia (la mente) y la materia física son sustancias (o modos de existencia) separadas, lo que da pie a que se plantee el peliagudo problema de cómo pueden llegar a interactuar la una con la otra. Hoy en día, pocos filósofos o científicos se adherirían

explícitamente a esta perspectiva. Pero el dualismo sigue seduciendo a otras muchas personas, al menos en Occidente. La intuitiva –y tan seductora– idea de que las experiencias conscientes *no parecen* físicas alienta un «dualismo ingenuo» en el que dicha «apariencia» alimenta ciertas creencias acerca de cómo son las cosas en realidad. Pero, como veremos a lo largo de este libro, lo que las cosas parecen suele ser muy mala guía de cara a saber cómo son de verdad.

Una variante particularmente influyente del fisicalismo es el *funcionalismo,* que, como aquel, es un supuesto muy común y, a menudo, no explícito de muchos neurocientíficos. Muchos de los que dan el fisicalismo por sentado también hacen lo propio con el funcionalismo.[10] Mi propia posición al respecto, sin embargo, es de un agnosticismo ligeramente suspicaz.

El funcionalismo es la idea según la cual la conciencia no depende de aquello de lo que un sistema está hecho (de su constitución física), sino solo de lo que el sistema hace, es decir, de las funciones que desempeña, de cómo transforma *inputs* en *outputs.* La intuición en la que se apoya el funcionalismo es que la mente y la conciencia son formas de procesamiento de información que los cerebros pueden poner en práctica, pero para las que estos (los cerebros biológicos) no resultan estrictamente necesarios.

Fíjate en cómo el concepto «procesamiento de información» se ha colado aquí sin avisar (como también lo hizo en la cita de Chalmers de unas páginas más atrás). Este concepto prevalece hasta tal punto en los debates y análisis sobre la mente, el cerebro y la conciencia que es fácil que se deslice dentro de cualquiera de ellos. Pero esto sería un error, porque la insinuación de que el cerebro «procesa información» oculta ciertos supuestos de hondo calado. Dependiendo de quién sea el que suponga, esas presuposiciones pueden ir desde la idea de que el cerebro es una especie de ordenador, en el que la mente (y la conciencia) serían el *software* (o *mindware),* hasta ciertas suposiciones sobre lo que la información en sí *es* realmente. Todos estos supuestos son peligrosos. Los

cerebros difieren sustancialmente de los ordenadores o, cuando menos, del tipo de ordenadores con el que estamos familiarizados.[11] Y la pregunta de qué *es* la información es casi igual de enojosa que la de qué es la conciencia, como veremos más adelante en este mismo libro. Esos problemas son la razón por la que recelo del funcionalismo.

Tomarse el funcionalismo al pie de la letra, como hacen muchos, conlleva la sorprendente implicación de que la conciencia sea algo que se puede *simular* en un ordenador. Recordemos que, para los funcionalistas, la conciencia solo depende de qué hace un sistema, no de qué está hecho. Eso quiere decir que, si logras entender correctamente las relaciones funcionales –si te aseguras de que un sistema cuenta con los diagramas de *inputs-outputs* correctos–, ya tendrás todo lo necesario para originar la conciencia. Dicho de otro modo, para los funcionalistas, *simulación* significa *sustanciación:* es como si se materializara en la realidad.

¿Hasta qué punto es razonable algo así? Para algunas cosas, la simulación sí equivale a sustanciación. Un ordenador que juega al go y es el mejor del mundo en ello, como el AlphaGo Zero de la empresa británica de inteligencia artificial DeepMind, *está jugando realmente al go.*[12] Pero hay muchas situaciones en las que eso no es así. Pensemos en los pronósticos del tiempo. A las simulaciones informáticas de los sistemas meteorológicos, por muy detalladas que sean, ni las moja la lluvia ni las zarandea el viento. ¿Se parece entonces la conciencia más al go o al viento? No esperes una respuesta, porque no la hay, al menos no todavía. Basta con que nos demos cuenta de que esa es una pregunta muy válida.[13] Por eso soy agnóstico ante el funcionalismo.

Quedan dos «ismos» más y ya habremos acabado.

El primero es el *pampsiquismo.* El pampsiquismo es la idea según la cual la conciencia es una propiedad tan fundamental del universo como otras propiedades fundamentales (como la masa-energía y la carga, por ejemplo) y está presente hasta cierto grado en todo y en todas partes. La gente se ríe a veces del pampsiquismo por afirmar que cosas como las

piedras y las cucharas son conscientes en el mismo sentido en que tú o yo lo somos, pero estas suelen ser malinterpretaciones deliberadas dirigidas a hacerlo quedar mal. Hay otras versiones más sofisticadas de esa misma idea y nos iremos reencontrando con algunas de ellas en capítulos posteriores del libro, pero los problemas principales del pampsiquismo no residen en su presunta insensatez (después de todo, siempre ha habido ideas descabelladas que han terminado por resultar certeras o, cuando menos, útiles). Sus problemas principales son que no explica nada en realidad y que no produce hipótesis falsables. Representa una salida fácil al aparente misterio planteado por el problema difícil que, una vez tomada, conduce a la ciencia de la conciencia hacia una vía muerta en lo empírico.[14]

Por último, está el *misterianismo,* habitualmente asociado al filósofo Colin McGinn.[15] El misterianismo es la idea según la cual, si bien podría existir una explicación física integral de la conciencia –una solución completa al problema difícil de Chalmers–, los seres humanos no somos lo bastante inteligentes (ni lo seremos nunca) para descubrirla o, siquiera, para reconocerla aunque unos extraterrestres superinteligentes nos la regalaran. Hay una explicación física de la conciencia, sí, pero está tan lejos de nosotros como una explicación de las criptomonedas pueda estarlo de las ranas. Nos está cognitivamente vedada por las limitaciones mentales propias de nuestra especie.

¿Qué podemos decir del misterianismo? Es muy posible que haya cosas que jamás comprendamos debido a las limitaciones de nuestros cerebros y mentes. De hecho, ya ocurre que nadie es capaz de comprender del todo cómo funciona un Airbus A380 (y, sin embargo, me encanta viajar en él, como ya hice en una ocasión, en un viaje de vuelta a casa desde Dubái). Hay, sin duda, cosas que siguen siendo cognitivamente inaccesibles para la mayoría de nosotros, aun cuando sean comprensibles en principio para los humanos, como los detalles más concretos de la teoría de cuerdas en física. Como los cerebros son sistemas físicos con recursos finitos, y como, según parece, algunos

cerebros son incapaces de comprender ciertas cosas, cabe suponer que hay cosas que son como son, pero que ningún humano podrá llegar nunca a conocer bien.[16] No obstante, es de un pesimismo injustificable presuponer que la conciencia se encuentra dentro de ese inexplorable terreno de la ignorancia inherente a nuestra especie.

Uno de los aspectos más maravillosos del método científico es su carácter acumulativo y progresivo. Hoy en día, muchos podemos comprender cosas que habrían resultado del todo incomprensibles –*incluso en principio*– para nuestros antepasados, y tal vez hasta para científicos y filósofos de hace apenas unas décadas. Con el tiempo, ha ido cayendo un misterio tras otro merced a la aplicación sistemática de la razón lógica y la experimentación. Si nos tomáramos el misterianismo en serio, ¿qué seguimos haciendo aquí? Más nos valdría recoger ya nuestros bártulos e irnos a casa. Así que mejor no le demos ninguna credibilidad.

Estos «ismos» representan diferentes formas de concebir la relación entre la conciencia y el universo en su conjunto. A la hora de sopesar sus respectivos méritos y deméritos, es importante que nos demos cuenta de que lo que más importa no es cuál de esos marcos es el «correcto» (en el sentido de cuál es probablemente verdad), sino cuál es más útil para ayudarnos a avanzar en nuestro conocimiento de la conciencia. Por eso tiendo a un tipo de fisicalismo agnóstico en el aspecto funcional. Para mí, esta es la mentalidad más pragmática y productiva que podemos adoptar cuando tratamos de desarrollar una ciencia de la conciencia. Es, además, en lo que a mí respecta, la más honesta desde el punto de vista intelectual.[17]

Pese a su atractivo, el fisicalismo no goza ni mucho menos de una aceptación universal entre los investigadores de la conciencia. Uno de los cuestionamientos más comunes que de él se hacen se basa en el llamado experimento mental «de los zombis». Los muertos vivientes en cuestión no son los semicadáveres devoradores de cerebros que aparecen en las películas, sino unos «zombis filosóficos». Pero son criaturas de las

que necesitamos deshacernos igualmente, porque, si no, la posibilidad de una explicación natural (fisicalista) de la conciencia habrá muerto antes incluso de nacer.

Un zombi filosófico es un ser indistinguible de una criatura consciente, salvo por el hecho de que carece de conciencia. Un Anil Seth zombi tendría mi aspecto, se comportaría como yo, andaría como yo y hablaría como yo, pero no tendría un modo de *ser* él, ni ningún universo interior, ni ninguna experiencia sentida. Si le preguntases al Anil zombi si está consciente, él te diría: «Sí, estoy consciente». El Anil zombi incluso habría escrito varios artículos y ensayos sobre la neurociencia de la conciencia, incluyendo ideas sobre la cuestionable relevancia de los zombis filosóficos para este tema. Pero nada de todo eso estaría implicando experiencia consciente alguna.[18]

Pues bien, se supone que la idea de los zombis proporciona un argumento desacreditador de las explicaciones fisicalistas de la conciencia porque, si somos capaces de imaginarnos un zombi, esto quiere decir que podemos concebir un mundo que sea indistinguible del nuestro, pero en el que no se esté produciendo conciencia alguna. Y si podemos concebir un mundo así, entonces es imposible que la conciencia sea un fenómeno físico.

Pero he aquí la razón por la que esa idea no es válida. El argumento de los zombis, como otros muchos experimentos mentales que tratan de desacreditar el fisicalismo, es un argumento sobre lo concebible, y los argumentos sobre lo concebible son intrínsecamente débiles. Como muchos de esos argumentos, posee una verosimilitud inversamente proporcional a la cantidad de conocimiento que se tiene al respecto.

¿Puedes imaginarte un A380 volando hacia atrás? Por supuesto que sí. Basta con que imagines un avión muy grande en el aire moviéndose hacia atrás. ¿Pero resulta *realmente* concebible una situación así? Pues, en realidad, cuanto más sabes de aerodinámica y de ingeniería aeronáutica, menos concebible te resulta. En este caso, basta con unos conocimientos mínimos sobre estos temas para que tengas muy claro

que los aviones no pueden volar marcha atrás. Sencillamente, *no es factible.*[*]

Lo mismo ocurre con los zombis. En cierto sentido, es trivial imaginarse un zombi filosófico. Solo tengo que imaginarme una versión de mí mismo deambulando por ahí sin tener experiencias conscientes. Pero ¿puedo *realmente* concebir algo así? Lo que se me está pidiendo que haga, *en realidad,* es considerar las capacidades y las limitaciones de una ingente red de muchos miles de millones de neuronas y de infinidad de sinapsis (las conexiones interneuronales), por no hablar de las células gliales y los gradientes de neurotransmisores, entre otras exquisiteces neurobiológicas, todo ello empaquetado dentro de un cuerpo que interactúa con un mundo en el que están incluidos otros cerebros inscritos en otros cuerpos.[19] ¿Puedo hacer algo así? ¿Puede hacerlo alguien? Lo dudo. Como en el caso del A380, cuanto más sabemos sobre el cerebro y su relación con las experiencias y las conductas conscientes, menos concebible resulta la posibilidad del zombi.[*]

El hecho de que algo sea o no concebible suele obedecer a una apreciación psicológica sobre la persona que concibe, no a una averiguación sobre la naturaleza de la realidad. Ese es el punto débil de los zombis. Se nos pide que imaginemos lo inimaginable, y que, de ese acto de comprensión ilusoria, extraigamos una serie de conclusiones acerca de los límites de la explicación fisicalista.[20]

Ya estamos listos, pues, para enfrentarnos a lo que yo llamo el *problema real* de la conciencia. Se trata de una forma de concebir la ciencia de la conciencia que, en mi caso, ha ido cobrando cuerpo a lo largo de muchos años, durante los que he ido asimilando las ideas de otras muchas personas y tomándolas como fundamento.[21] Abordar el problema real es, a mi entender, el enfoque mediante el que más probabilidades de éxito podrá tener una ciencia de la conciencia.

Cuando se enfoca el problema real, los objetivos primordiales de la ciencia de la conciencia son *explicar, predecir* y *controlar* las propiedades fenomenológicas de la experiencia consciente. Eso significa explicar por

qué una experiencia consciente concreta es como es –por qué posee las propiedades fenomenológicas que posee– en términos de mecanismos físicos y de procesos en el cerebro y el organismo. Gracias a esas explicaciones, deberíamos poder predecir cuándo ocurrirán experiencias subjetivas específicas, y también deberíamos poder controlarlas interviniendo en los mecanismos subyacentes. En resumen, abordar el problema real exige explicar *por qué* un patrón particular de actividad cerebral –u otros procesos físicos– se corresponde con un tipo particular de experiencia consciente, y no limitarse a demostrar que lo hace.

El problema real es distinto del problema difícil, ya que no gira (o no de entrada, al menos) en torno a la pregunta de por qué y cómo es que hay conciencia y forma parte del universo. No se trata de dar con ese algo especial que puede aportarle la magia a lo que, en principio, es mero mecanismo. También difiere del problema fácil (o problemas fáciles), porque se centra en la fenomenología en vez de en las funciones o los comportamientos. No barre bajo la alfombra los aspectos subjetivos de la conciencia. Y por su énfasis en los mecanismos y los procesos, el problema real concuerda de manera natural con una cosmovisión fisicalista sobre la relación entre materia y mente.

Para aclarar estas distinciones, preguntémonos cómo trataría de explicar cada uno de esos diferentes enfoques nuestra experiencia subjetiva de la «rojez».

Desde una perspectiva de problema fácil, el reto consistiría en explicar todas las propiedades mecánicas, funcionales y conductuales asociadas al hecho de experimentar la rojez: el cómo unas longitudes de onda luminosa específicas activan el sistema visual, las condiciones en las que decimos cosas como que «ese objeto es rojo», el comportamiento típico ante los semáforos, el cómo las cosas rojas inducen a veces respuestas emocionales de un carácter determinado, etcétera.

El enfoque del problema fácil se abstiene –con toda la intención– de abordar cualquier explicación de por qué y cómo esas propiedades funcionales, mecánicas y conductuales deberían acompañarse de

fenomenología alguna (en este caso, de la fenomenología de la «rojez»). La existencia de una experiencia subjetiva –por oposición a una no experiencia– es ya terreno propio del problema difícil. Por mucha información que te den sobre los mecanismos, nunca dejará de ser razonable que te preguntes: «De acuerdo, pero ¿por qué este mecanismo está asociado con la experiencia consciente?». Si te tomas en serio el problema difícil, siempre sospecharás de la existencia de una brecha explicativa entre las explicaciones mecánicas y la experiencia subjetiva de «ver rojo».

El problema real, por su parte, acepta la existencia de las experiencias conscientes y se enfoca principalmente en sus propiedades fenomenológicas, como, por ejemplo, que las experiencias de la rojez sean visuales, que habitualmente (aunque no siempre) se asocien a objetos, que parezcan ser propiedades de las superficies, que muestren diferentes niveles de saturación, que definan una categoría entre otras experiencias cromáticas (aunque puede haber variaciones sutiles dentro de esa misma categoría), etcétera. Lo importante es que todas estas son propiedades *de la experiencia en sí,* y no (o, por lo menos, no principalmente) de las funciones o comportamientos asociados a esa experiencia. El reto desde la perspectiva del problema real es explicar, predecir y controlar esas propiedades fenomenológicas en términos de cosas que ocurren en el cerebro y el organismo. Querríamos saber qué puede haber en los patrones específicos de actividad en el cerebro –como los complejos bucles de retroalimentación en el córtex visual–[*] que explique (y prediga, y controle) por qué una experiencia como la de la rojez es del modo particular en que es y no de otro. ¿Por qué no es como la experiencia de lo azul, o como un dolor de muelas, o como los celos?

Explicación, predicción y control. Estos son los criterios con arreglo a los que se valoran la mayoría de los demás proyectos científicos, con independencia de lo confusos que puedan parecer inicialmente los fenómenos que son sus objetos de estudio. Los físicos han dado pasos

grandísimos para desentrañar los secretos del universo –es decir, para explicar, predecir y controlar sus propiedades–, pero siguen estando muy desconcertados en lo que se refiere a determinar de qué está hecho o por qué existe siquiera. La ciencia de la conciencia también puede realizar grandes progresos de cara a arrojar luz sobre las propiedades y la naturaleza de las experiencias conscientes sin necesidad de explicar cómo ni por qué forman parte del universo en el que vivimos.

Tampoco deberíamos esperar necesariamente que las explicaciones científicas siempre sean intuitivamente satisfactorias. En física, es bien conocido el carácter contraintuitivo de la mecánica cuántica, que, pese a ello, goza de amplia aceptación como mejor explicación disponible de la naturaleza de la realidad física. Bien podría ocurrir que, en el futuro, una ciencia ya madura de la conciencia nos permita explicar, predecir y controlar propiedades fenomenológicas sin que nos llegue a proporcionar la intuitiva impresión de que «sí, seguro que es así, *¡es obvio* que es así!».

Es importante entender que centrarse en el problema real de la conciencia no representa admitir una derrota ante el problema difícil. De hecho, por indirectamente que sea, el problema real viene *después del* problema difícil. Para entender por qué, permíteme que te presente los «correlatos neurales de la conciencia».

Todavía me asombro de la mala reputación de la que gozaba la ciencia de la conciencia hace apenas treinta años. En 1989, un año antes de que empezara mis estudios de grado en la Universidad de Cambridge, el destacado psicólogo Stuart Sutherland escribió: «La conciencia es un problema tan fascinante como esquivo. Es imposible especificar qué es, qué hace o por qué surgió durante la evolución. Nada se ha escrito sobre ella que sea digno de ser leído».[22] Tan desdoroso resumen apareció nada menos que en un diccionario internacional de psicología así titulado *(The International Dictionary of Psychology),* y capta muy bien la actitud ante la cuestión de la conciencia con la que a menudo me encontré al dar mis primeros pasos en el mundo académico.

En otras latitudes, lejos de Cambridge, la situación era más prometedora, aunque yo no lo supiera por aquel entonces. Francis Crick (codescubridor, junto a Rosalind Franklin y James Watson, de la estructura molecular del ADN) y su colega Christof Koch, miembros ambos del mismo instituto de investigación en San Diego (California), estaban desarrollando el que se convertiría en el método dominante durante el auge que la ciencia de la conciencia experimentó a partir de entonces: me refiero a la búsqueda de los correlatos neurales de la conciencia.[23]

Según su definición estándar, un correlato neural de la conciencia (o CNC) es «el conjunto de los mecanismos neuronales mínimamente suficientes para que se produzca un percepto consciente específico cualquiera».[24] El postulado de partida del enfoque de los CNC es que para toda experiencia (por ejemplo, para la experiencia de «ver rojo») existe un patrón concreto de actividad neuronal responsable de la misma. Siempre que esta actividad esté presente, se producirá una experiencia de la rojez, y si aquella no lo está, esta no se producirá.

La gran virtud del enfoque de los CNC es que ofrece una receta práctica para proceder a la investigación. Para identificar un CNC, basta con que preparemos una situación en la que las personas tengan a veces una experiencia consciente particular y no la tengan otras veces, y con que nos aseguremos al mismo tiempo de que el resto de condiciones sean siempre lo más parecidas posible. Se trata entonces de comparar la diferencia de actividad en el cerebro de la persona entre la condición en la que se da la experiencia consciente y la condición en la que no, utilizando para ello técnicas de imagen cerebral como la imagen por resonancia magnética funcional (IRMf) o la electroencefalografía (EEG).[*] La actividad cerebral específica de la condición «consciente» refleja el CNC correspondiente a esa experiencia particular.[25]

Los estudios con «rivalidad binocular» son un buen ejemplo. La rivalidad binocular se induce mostrándole simultáneamente una imagen distinta a cada ojo: puede ser una foto de una cara al ojo izquierdo, y

otra de una casa al derecho. En esa situación, la percepción consciente no se queda en un punto medio, mostrándonos una grotesca quimera de cara y casa, sino que oscila entre la primera y la segunda imagen, deteniéndose durante unos segundos en cada una. Primero ves una casa, luego una cara, luego una casa de nuevo… y así sucesivamente. Lo importante en este caso es que la percepción consciente cambia a pesar de que el *input* sensorial se mantiene constante. Así pues, observando lo que ocurre en el cerebro, resulta posible distinguir entre la actividad cerebral que sigue el rastro de la percepción consciente y aquella otra que sigue el del *input* sensorial, sea cual sea este. La actividad cerebral que acompaña a la percepción consciente es la que nos identifica el CNC de dicha percepción.[26]

La estrategia de los CNC lleva ya muchos años mostrando una extraordinaria fertilidad y nos ha dado volúmenes y más volúmenes de hallazgos fascinantes, pero sus limitaciones empiezan a hacerse evidentes. Uno de sus problemas es que cuesta mucho –y puede que, en última instancia, sea imposible– desglosar un «verdadero» CNC de un conjunto de factores potencialmente confundidores, entre los que los más importantes son aquellos sucesos neurales que, o bien son prerrequisitos de un CNC propiamente dicho, o bien son consecuencias del mismo.[27] En el caso de la rivalidad binocular, por ejemplo, la actividad cerebral que acompaña a la percepción consciente puede también corresponderse con procesos «corriente arriba» (precondiciones) como el de «prestar atención», o con otros «corriente abajo» (efectos) como la conducta verbal de «informar», es decir, de contar que estás viendo una casa o una cara. Aunque relacionados con el flujo de la percepción consciente, no deberíamos confundir los mecanismos neurales responsables de la atención y del acto de informar verbalmente –o de otras precondiciones y consecuencias– con los que son responsables de la percepción consciente propiamente dicha.[28]

Otro de sus problemas, más profundo todavía, es que las *correlaciones* no son *explicaciones.* Todos sabemos que con una mera correlación no se

puede establecer una causación, pero no es menos cierto que tampoco alcanza para una explicación. Ni siquiera con diseños experimentales cada vez más ingeniosos y tecnologías de imagen cerebral más y más potentes, se puede conseguir que la correlación por sí sola alcance a ser una explicación. Vistos desde esa perspectiva, la estrategia de los CNC y el enfoque del problema difícil son dos compañeros de cama naturales. Si nos ceñimos a reunir correlaciones entre cosas que suceden en el cerebro y cosas que suceden en nuestra experiencia, no es de extrañar que sospechemos siempre de la existencia de una brecha explicativa entre lo físico y lo fenoménico. Pero, en cambio, si vamos más allá del establecimiento de correlaciones y descubrimos explicaciones que conectan propiedades de los mecanismos neuronales con propiedades de la experiencia subjetiva –como se defiende desde el enfoque del problema real–, esa brecha se irá cerrando y puede que incluso llegue a desaparecer por completo. Cuando seamos capaces de predecir (y explicar, y controlar) por qué la experiencia de la rojez es del modo particular en que es –y no como la de lo azul, o como la de los celos–, el misterio de cómo surge la rojez será menos misterioso o, tal vez, dejará de serlo del todo.

La aspiración última del enfoque del problema real es que, a medida que vayamos construyendo puentes explicativos cada vez más robustos entre lo físico y lo fenomenológico, esa idea intuitiva del problema difícil de que jamás se podrá comprender la conciencia en términos físicos vaya perdiendo su razón de ser y termine por disiparse entre una nube de humo metafísico. Cuando lo haga, tendremos en nuestras manos una ciencia de la experiencia consciente tan provechosa como plenamente satisfactoria.[29]

¿Qué hay que justifique semejante aspiración? Pensemos en cómo ha ido madurando el conocimiento científico de la *vida* durante el último par de siglos.

No hace tanto que la vida parecía tan misteriosa como la conciencia lo parece hoy. Los científicos y los filósofos de entonces dudaban de que

los mecanismos físicos o químicos pudieran explicar nunca la propiedad de *estar vivo*. La diferencia entre lo viviente y lo no viviente, o entre lo animado y lo inanimado, parecía tan fundamental que se juzgaba harto inverosímil que jamás pudiese ser salvada por explicación mecánica alguna.

Esta filosofía del *vitalismo* llegó a su apogeo en el siglo XIX. La apoyaban destacados biólogos como Johannes Müller y Louis Pasteur, y pervivió hasta bien entrado el siglo XX. Los vitalistas creían que la propiedad de estar vivo solo se podía explicar apelando a un algo especial: una chispa de vida, un *élan vital.*[30] Pero como hoy bien sabemos, no se necesita ningún elemento especial. En nuestros días, el vitalismo está absolutamente desacreditado en los círculos científicos. Aunque sigue habiendo muchas cosas que desconocemos acerca de la vida –cómo funciona una célula, por ejemplo–, la idea de que estar vivo requiere de cierto ingrediente sobrenatural ha perdido toda credibilidad. El error fatal del vitalismo consistió en confundir una falta de imaginación con la intuición de una necesidad. Se trata del mismo error que encontramos en la raíz del argumento de los zombis.

La ciencia de la vida pudo superar la miopía del vitalismo gracias a su focalización en el progreso práctico, es decir, al énfasis en los «problemas reales» de lo que significa estar vivo. Sin dejarse intimidar por el pesimismo vitalista, los biólogos prosiguieron con la labor de describir las propiedades de los sistemas vivos y, luego, explicar (y también predecir y controlar) cada una de esas propiedades en términos de mecanismos físicos y químicos. La reproducción, el metabolismo, el crecimiento, la autorreparación, el desarrollo, la autorregulación homeostática: todo ello terminó abriéndose individual y colectivamente a las explicaciones mecánicas. A medida que se fueron rellenando los detalles (que todavía se siguen completando), no solo se disipó el misterio fundamental de «qué es la vida», sino que el propio concepto de vida se ramificó, de tal modo que «estar vivo» es algo que ya no se concibe como una sola propiedad de todo o nada. Fueron surgiendo

áreas grises; conocido fue el caso de los virus, pero ahora también lo es el de los organismos sintéticos e incluso de los cúmulos de gotículas de aceite que poseen algunas, aunque no todas, las propiedades características de los sistemas vivos. La vida se naturalizó y se volvió más fascinante si cabe al hacerlo.

Este paralelismo nos proporciona tanto un motivo de optimismo como una estrategia práctica para abordar el problema real de la conciencia.

El optimismo reside en que los investigadores actuales de la conciencia podrían hallarse en una situación similar a la que afrontaban los biólogos cuando estudiaban la naturaleza de la vida hace solo unas pocas generaciones. Lo que se considera misterioso *ahora* podría no serlo *siempre*. A medida que avancemos en la explicación de las diversas propiedades de la conciencia en términos de sus mecanismos subyacentes, es posible que el misterio fundamental de «cómo sucede la conciencia» termine por disolverse igual que el misterio de «qué es la vida» se disolvió en su momento.

Lógicamente, el paralelismo entre la vida y la conciencia no es perfecto. La diferencia más visible es que las propiedades de la vida resultan objetivamente descriptibles, mientras que las dianas explicativas a las que apunta la ciencia de la conciencia son subjetivas y solo existen en primera persona. No obstante, no se trata de ninguna barrera infranqueable; solo significa, básicamente, que los datos relevantes son más difíciles de recopilar porque son subjetivos.

La estrategia práctica se fundamenta en la idea de que la conciencia, como la vida, no es un fenómeno único. Cuando los biólogos apartaron su foco de atención de la vida en general como si esta fuera un único y temible gran misterio, dejaron de buscar (o necesitar) con tanto ahínco hallar una solución maravillosa, un momento «eureka», y pasaron a dividir el «problema» de la vida en una serie de procesos relacionados entre sí, pero diferenciables.[31] Aplicando esa misma estrategia a la conciencia, en este libro me centraré por separado en el *nivel,* el *contenido* y el *yo* como propiedades centrales de lo que significa *ser tú.* Al

hacerlo, se nos revelará una imagen satisfactoria del conjunto de la experiencia consciente.

El *nivel* de la conciencia hace referencia a «cuán conscientes somos» en una escala que va desde la ausencia absoluta de toda experiencia consciente (como ocurre en los estados de coma o de muerte cerebral) hasta los estados de consciencia clara e intensa que acompañan a nuestra vida despierta normal.

El *contenido* de la conciencia tiene que ver con aquello de lo que somos conscientes: las visiones, sonidos, olores, emociones, estados de ánimo, pensamientos y creencias que forman nuestro universo interior. Los contenidos conscientes son todos ellos variedades de *percepción,* es decir, interpretaciones de base cerebral de unas señales sensoriales que conforman colectivamente nuestras experiencias conscientes. (La percepción, como veremos, puede ser tanto consciente como inconsciente).

Por último, está el *yo* consciente: la experiencia concreta de *ser tú,* que es el tema conductor del presente libro. La experiencia de «ser un yo» es un subconjunto de contenidos conscientes que comprende experiencias de tener un cuerpo determinado, una perspectiva de primera persona y un conjunto de recuerdos únicos, además de experiencias de estados de ánimo, de emociones y del famoso «libre albedrío». La yoidad es probablemente el aspecto de la conciencia al que más nos aferramos, tanto que podemos sentirnos tentados a confundir la autoconciencia (la experiencia de ser un yo) con la conciencia en sí (la presencia de cualquier clase de experiencia subjetiva, de cualquier fenomenología).

Al hacer estas distinciones, no estoy diciendo que esos aspectos de la conciencia sean completamente independientes unos de otros. De hecho, no lo son, y averiguar cómo se interrelacionan representa otro importante reto para la ciencia de la conciencia.

No obstante, subdividir el problema real de la conciencia en esos amplios términos presenta numerosas ventajas. Cuando se señalan objetivos diferenciados para la explicación, resulta más viable proponer

posibles mecanismos capaces de acometer las tareas requeridas de explicación, predicción y control. No menos importante es el hecho de que la subdivisión ayuda a combatir la limitadora idea de que la conciencia no es más que «una cosa», un único e intimidante misterio que tal vez escape por completo a la explicación científica. Sin embargo, aquí veremos cómo esas diferentes propiedades de la conciencia se combinan de modos diversos, cruzando fronteras entre especies e incluso entre personas distintas. Existen tan numerosas formas distintas de estar consciente como organismos conscientes desemejantes hay.

Al final, es posible que el propio problema difícil sucumba y que podamos por fin comprender la continuidad entre la conciencia y el resto de la naturaleza sin necesidad de adoptar ningún «ismo» arbitrario que «decrete» de qué manera se relacionan la fenomenología y la física.

De hecho, esa es la promesa que nos hace el enfoque del problema real. Si quieres ver cuán lejos nos lleva, sigue leyendo.

2. MEDIR LA CONCIENCIA

¿Cuán consciente estás ahora mismo? ¿Qué marca la diferencia entre estar consciente y ser un simple pedazo de carne viviente –o de silicio inerte– sin universo interior alguno? Una serie de nuevas teorías y tecnologías están permitiendo a los científicos, por vez primera, medir *niveles* de conciencia. Para comprender mejor estas nuevas investigaciones, fijémonos en las raíces de su desarrollo.

En el París del siglo XVII, había un sótano profundo, oscuro y frío bajo el Observatoire, en la margen izquierda del Sena. Este subterráneo desempeñó un sorprendente papel en la historia de la ciencia: uno que pone muy de relieve la importancia de la *medición* en el avance del saber.

Varios filósofos y científicos de la época –aunque no se les llamaba aún así, científicos– libraban una especie de competición por ser los primeros en desarrollar termómetros fiables y, de paso, lograr un mejor conocimiento físico de la naturaleza del calor. Las teóricas «caloríficas» populares, según las cuales el calor era una sustancia que podía fluir entrando en las cosas o saliendo de ellas, estaban cayendo en desuso. La revisión de esas teorías dependía de unos experimentos precisos en los que se pudiera calibrar de forma sistemática la «calidez» o la «frialdad» de los objetos. Eran experimentos que precisaban tanto de un método de medición del «calor», fuera lo que fuere este, como de una escala con la que comparar las diferentes mediciones. Había una carrera en marcha, pues, por desarrollar un *termómetro* y una escala de *temperatura* fiables. Pero ¿cómo se podría establecer la fiabilidad de un termómetro si antes

no se contaba con una escala contrastada con la que validar aquella? ¿Y cómo desarrollar una escala de temperatura sin disponer antes de un termómetro fiable?

El primer paso hacia la resolución de este dilema era dar con un punto fijo de partida: una referencia invariable de la que se pudiera suponer que mantenía una temperatura constante. Pero incluso esto suponía todo un reto. Se sabía que algunos candidatos prometedores, como el punto de ebullición del agua, dependían de factores como la presión atmosférica, que variaba con la altitud y la meteorología, e incluso con influencias tan sutiles como la rugosidad de la superficie de un vaso de agua. Fue debido a frustraciones de ese tipo por lo que, durante un tiempo, se pensó que un sótano parisino –con su frescor aparentemente constante– sería una opción razonable para establecer ese punto fijo de temperatura. (No fue la única sugerencia extraña. Quizá la palma en ese sentido se la llevase un tal Joachim Dalencé, quien propuso como referencia el punto de fusión de la mantequilla).[1]

Al final, se inventaron termómetros de mercurio, fiables y precisos, que posibilitaron dejar atrás las teorías caloríficas y fundar una nueva ciencia de la termodinámica, toda una revolución asociada a figuras legendarias como Ludwig Boltzmann y Lord Kelvin. En la termodinámica, la temperatura es una propiedad a gran escala del movimiento de las moléculas dentro de una sustancia, concretamente, de la energía cinética molecular media. A movimiento más rápido, mayor temperatura. El «calor» pasa así a ser la energía transferida entre dos sistemas que están a temperaturas diferentes. Es importante entender que la termodinámica hizo algo más importante que el simple hecho de establecer que la energía cinética media está *correlacionada* con la temperatura, pues propuso que la temperatura *es* eso mismo en realidad. Equipados con esta nueva teoría, los científicos pudieron hablar ya de cosas como la temperatura de la superficie del sol, e incluso identificaron un «cero absoluto», en el que todo movimiento molecular cesa (al menos, en teoría). A partir de ahí, se sustituyeron otras escalas

anteriores, basadas en mediciones de sustancias particulares (como la Celsius o la Fahrenheit) por otra basada en propiedades físicas subyacentes (la Kelvin, así nombrada en honor del lord epónimo). La base física de la temperatura y el calor ya no es ningún misterio.

Leí por vez primera esta historia en el libro *Inventing Temperature,* del historiador Hasok Chang, profesor del University College de Londres.[2] Hasta entonces, nunca había sabido apreciar del todo hasta qué punto el progreso científico depende de la medición. La historia de la termometría y su repercusión en nuestra concepción del calor representan un ejemplo manifiesto de cómo la capacidad de obtener unas medidas cuantitativas detalladas, sobre una escala definida por puntos de referencia fijos, tiene en sí misma el potencial de transformar algo misterioso en algo comprensible.

¿Podría funcionar ese mismo enfoque para la conciencia?[3]
Los filósofos hablan a veces de un hipotético «medidor de la conciencia» que sería capaz de determinar si algo (otra persona, animal o, tal vez, incluso máquina) es un ser consciente o no. En una conferencia en los años noventa, en pleno apogeo del enfoque del problema difícil, David Chalmers cogió un viejo secador de mano y se apuntó con él a la cabeza para dejar claro lo útil que un aparato así sería si existiera. Apuntas con tu medidor de conciencia a algo y miras qué lectura te da: se acabaron los misterios sobre hasta dónde se extiende el círculo mágico de la conciencia.

Sin embargo, tal como nos muestra la historia de la temperatura, el valor de la medición radica, no ya en que nos proporcione respuestas de «sí o no» acerca de la presencia o ausencia de una propiedad, sino en que posibilite experimentos cuantitativos que posean el potencial de transformar el conocimiento científico del fenómeno en cuestión.[4]

Si la conciencia acaba siendo algo análogo a la temperatura (es decir, si se trata de un proceso físico único, subyacente e idéntico al hecho de «estar consciente»), los beneficios serían espectaculares. No solo podríamos determinar «cuán consciente» está alguien, sino que

podríamos hablar con sentido sobre «niveles» y «grados» específicos de conciencia, así como sobre variedades de conciencia muy alejadas del ejemplo humano, tan local nuestro.

Pero incluso en el caso de que la historia de la conciencia termine por ser distinta (menos parecida a la de la temperatura que a la de la vida, o al menos, así lo sospecho yo), la capacidad de obtener mediciones precisas no dejaría de constituir un paso esencial de cara a la construcción de puentes explicativos –que es como decir de cara a la explicación, la predicción y el control– de la naturaleza de las experiencias subjetivas. Tanto en un caso como en el otro, la medición transformaría lo cualitativo en cuantitativo, y lo vago en preciso.

La medición posee también unas motivaciones prácticas. El arte de la anestesia –aplicada a más de cuatro millones de personas al día–[5] implica mantener a un paciente en una inconsciencia temporal sin inducirle una sobredosis. Un medidor fiable y preciso de la conciencia tendría un valor evidente de cara a alcanzar ese delicado equilibrio, sobre todo porque la anestesia suele ir acompañada de unos agentes bloqueantes neuromusculares que inducen una parálisis temporal para que los cirujanos puedan realizar su trabajo sin tener que lidiar con incómodas reacciones musculares reflejas en las personas intervenidas. Y, como enseguida veremos, hay una necesidad urgente de encontrar nuevos métodos que nos permitan decidir si la conciencia se mantiene de algún modo en un paciente cuando sufre una lesión cerebral grave y se le diagnostican situaciones tan aterradoras como un «estado vegetativo» o un «estado de mínima conciencia».

Lo cierto es que hace ya muchos años que en los quirófanos se emplean monitores de conciencia cerebral. El más habitual es el monitor de «índice biespectral».[6] Aunque los detalles se mantienen guardados a buen recaudo bajo una capa de patentes, el concepto en el que se basan es el de combinar toda una serie de medidas electroencefalográficas (EEG) en un solo indicador que se va actualizando continuamente y que guía al anestesista durante la intervención quirúrgica. La idea es buena,

pero los monitores de índice biespectral siempre han estado rodeados de controversia, en parte, porque ha habido varios casos en los que sus lecturas no han concordado con otros signos conductuales de conciencia, como que los pacientes abrieran los ojos, o como que recordaran lo que los cirujanos estaban diciendo durante la operación.[7] Pero un problema más de fondo aún en lo que a la propia ciencia de la conciencia se refiere es que el índice biespectral no se basa en ninguna teoría derivada de principio alguno.

Durante los últimos años, una nueva generación de medidores de conciencia han empezado a cobrar forma para ser usados, no en quirófanos, sino en laboratorios neurocientíficos. A diferencia de otros monitores de conciencia anteriores, estos nuevos métodos están estrechamente vinculados a una comprensión teórica emergente de las bases cerebrales del estado consciente, y ya están comenzando a demostrar su valía práctica.

Medir el nivel consciente en humanos no es lo mismo que determinar si alguien está despierto o dormido. Nivel consciente y activación fisiológica *(arousal)* no son la misma cosa. Aunque ambos suelen estar muy correlacionados, la conciencia (o estado consciente) y la vigilia (o estado despierto o activación *[arousal])* pueden separarse la una de la otra en varios sentidos, y esto basta para mostrar que no pueden depender de la misma base biológica subyacente. Por ejemplo, y por definición, cuando sueñas, estás dormido, pero también estás teniendo experiencias conscientes, ricas y variadas. En el extremo opuesto, ubicaríamos las situaciones con diagnóstico catastrófico, como el estado vegetativo (también llamado «síndrome de vigilia sin respuesta»), en el que una persona todavía experimenta ciclos de sueño y vigilia, pero no muestra señales conductuales de un estado consciente: es como si en una casa se encendieran las luces de vez en cuando, pero no hubiera nadie dentro. La figura de la página siguiente ilustra la relación entre el estado consciente y el estado despierto en diversas situaciones, tanto normales como patológicas.

Para monitorizar el nivel consciente, necesitamos preguntarnos qué subyace en el cerebro de una persona al hecho de que esté consciente, a diferencia del hecho de que esté simplemente despierta. ¿Acaso es una cuestión del número de neuronas implicadas? No lo parece. El cerebelo (el «pequeño cerebro» que pende de la parte trasera de tu córtex) tiene aproximadamente el cuádruple de neuronas que el resto del encéfalo, pero apenas parece participar en la conciencia.[8] Existe un raro trastorno llamado agenesia cerebelosa por el que algunas personas no logran desarrollar un cerebelo normal y, aun así, se las arreglan bastante bien para llevar una vida corriente (desde luego, no dan pie alguno a sospechar que no estén conscientes).[9]

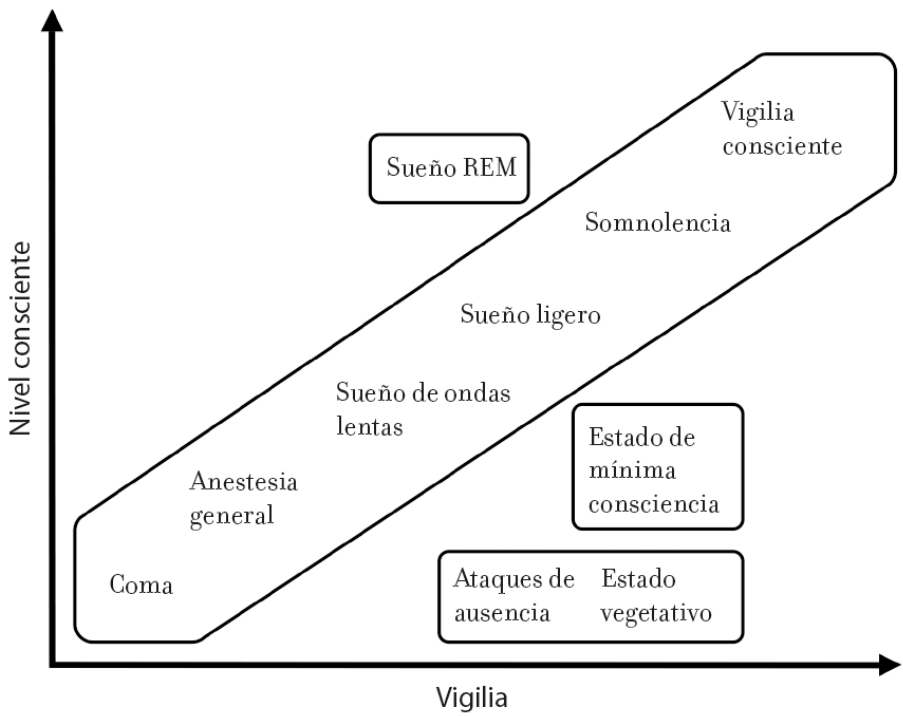

Fig. 1. La relación entre nivel consciente (consciencia) y vigilia (activación)

¿Y el nivel total de actividad neuronal? ¿Está más activo el cerebro, por lo general, durante los estados conscientes que durante los

inconscientes? Sí, tal vez, hasta cierto punto…, pero no por mucho. Aunque se aprecian diferencias en el consumo de energía del cerebro entre los diferentes niveles de conciencia, estas diferencias son bastante pequeñas y, desde luego, no existe sentido alguno en el que se pueda decir que el cerebro «se apaga» cuando la conciencia se desvanece.[10]

La conciencia parece depender, más bien, de cómo diferentes partes del cerebro se comunican entre sí. Y no me refiero al cerebro en su conjunto: los patrones de actividad que importan parecen ser los que se producen dentro del sistema talamocortical, que es la combinación que forman la corteza cerebral y el tálamo (un conjunto de estructuras cerebrales de forma ovalada –«núcleos»– situado justo debajo del córtex y muy estrechamente conectado con él). Los últimos y más interesantes enfoques que se están aplicando para medir el nivel consciente –y para distinguirlo del estado de vigilia– se basan en monitorizar y cuantificar esas interacciones. La versión más ambiciosa de esta idea produce incluso un único número que indica lo consciente que está una persona. Igual que un termómetro.

El pionero de la aplicación de este nuevo enfoque ha sido el neurocientífico italiano Marcello Massimini, inicialmente en colaboración con el reputado investigador de la conciencia Giulio Tononi en la Universidad de Wisconsin en Madison, y en época más reciente, con su propio grupo en la Universidad de Milán. Lo que hicieron fue simple y elegante. Para testar cómo diferentes partes de la corteza cerebral se hablaban entre sí, estimularon la actividad en un punto y registraron cómo ese pulso se extendía en el espacio y en el tiempo a otras regiones corticales. Lo hicieron combinando dos técnicas: la EEG y la estimulación magnética transcraneal (EMT). Un equipo de EMT consiste en un electroimán controlado con gran precisión que permite que el investigador inyecte un breve y agudo pulso de energía directamente en el cerebro atravesando el cráneo, mientras que la EEG se usa en este caso para registrar la respuesta del cerebro a esa irrupción. Es como si alguien sacudiera el cerebro con un martillo

eléctrico y se detuviera a escuchar el eco posterior.

Puede que te sorprenda saber que las personas que la reciben casi nunca se enteran de la descarga de EMT en sí, salvo que provoque algo evidente como un movimiento (cuando el imán se coloca por encima de la corteza motora, que controla las acciones) o un destello visual simple (un «fosfeno», que se puede producir cuando se activa la corteza visual). Y si la descarga causa un espasmo de los músculos de tu rostro o de tu cuero cabelludo, notarás el dolor. Pero, por lo general, la enorme perturbación ocasionada por la EMT en la actividad cerebral no genera ninguna alteración en la experiencia consciente. Tal vez esto no sea tan sorprendente. Solo demuestra que no somos conscientes de lo que hacen nuestras neuronas… ¿Y por qué íbamos a serlo?

Aunque no sintamos directamente los pulsos de la EMT, Massimini y Tononi descubrieron que se podían aprovechar sus ecos eléctricos para distinguir entre diferentes niveles de conciencia. En los estados inconscientes, como el de cuando dormimos sin soñar, o el de cuando estamos bajo anestesia general, esos ecos son muy simples. Hay una respuesta inicial fuerte en la parte del cerebro que ha recibido la descarga inicial, pero esa respuesta se desvanece rápido, como las ondas provocadas por una piedra al caer sobre una superficie de agua en calma. Pero durante los estados conscientes, la respuesta es muy distinta: hay un eco característico que se extiende ampliamente por la superficie cortical, y que desaparece y reaparece conforme a unos patrones complejos. La complejidad de estos, tanto en el tiempo como en el espacio, implica que hay diferentes partes del cerebro –y, en particular, del sistema talamocortical– que se intercomunican de maneras mucho más sofisticadas durante los estados conscientes que durante los inconscientes.[11]

Aunque la diferencia entre unos y otros estados suele ser fácil de apreciar simplemente observando bien los datos, lo verdaderamente apasionante de este trabajo es que nos indica que se puede cuantificar la complejidad del eco. Es posible asignarle un número con el que

especificar la *magnitud* de dicha complejidad. Es una metodología conocida en inglés como de *zap and zip:* se usa la EMT para «sacudir» *(zap)* la corteza cerebral, y se usa luego un algoritmo informático para «comprimir» *(zip)* la respuesta –el eco eléctrico– en una cifra unitaria.

El algoritmo empleado para «comprimir» es el mismo que se usa para los formatos *zip* de compresión de fotografías digitales en archivos de menor tamaño. Todo patrón, ya sea una foto de tus vacaciones de verano, ya sea un eco eléctrico que se despliega temporal y espacialmente por el cerebro, puede representarse como una secuencia de unos y ceros. Toda secuencia no aleatoria se puede traducir a una representación comprimida, una cadena mucho más corta de dígitos que se puede usar para regenerar por completo el original. La longitud de la representación comprimida más corta posible es lo que se llama la *complejidad algorítmica* de la secuencia. La complejidad algorítmica será mínima en el caso de una secuencia completamente predecible (por ejemplo, una que esté formada totalmente por unos o por ceros) y máxima cuando se trata de una absolutamente aleatoria, y alcanzará niveles intermedios en los casos de secuencias que contengan cierta cantidad de estructura predecible. El algoritmo de compresión *zip* –que calcula la denominada «complejidad Lempel-Ziv-Welch» (o «complejidad LZW» para abreviar)– es una manera muy popular de estimar la complejidad algorítmica de cualquier secuencia dada.[12]

Massimini y su equipo llamaron a su medida de los ecos registrados en sus experimentos *índice de complejidad perturbacional,* o ICP. Este se vale de la complejidad LZW para proporcionar un indicador de la complejidad algorítmica de la respuesta del cerebro a una perturbación (el pulso de EMT).

Los investigadores validaron inicialmente su indicador mostrando que los valores del ICP durante estados inconscientes como el del sueño sin sueños, o el de la anestesia general, eran confiablemente más bajos que durante un estado consciente de referencia (el del descanso despierto). Eso es alentador, pero el verdadero potencial del método del ICP estriba

en que define una escala continua, lo que permite establecer distinciones más detalladas. En un trascendental estudio de 2013, el equipo de Massimini midió los valores del ICP de un gran número de pacientes con lesiones cerebrales que padecían trastornos de la conciencia.[13] Los investigadores descubrieron que las magnitudes del ICP estaban muy estrechamente correlacionadas con los niveles de discapacidad (diagnosticados por neurólogos de forma independiente en todos los casos). Por ejemplo, las personas en estado vegetativo, en las que se supone que la conciencia ya no está presente, pese a que conserven el estado de vigilia, tenían puntuaciones de ICP más bajas que las que se encontraban en estados de mínima consciencia, en las que las señales conductuales de la conciencia van y vienen. Incluso pudieron trazar una línea divisoria entre valores del ICP indicativos de conciencia de aquellos otros que sugieren la ausencia de esta.

En mi grupo de investigación de la Universidad de Sussex, hemos estado trabajando con métodos similares para valorar el nivel consciente. Pero en vez de usar la EMT para inyectar pulsos de energía en la corteza cerebral, nos hemos dedicado a medir la complejidad algorítmica de la actividad cerebral continuada y natural («espontánea», podríamos llamarla). Podríamos describirlo como la «compresión» sin la «descarga». En una serie de estudios dirigidos por mi homólogo Adam Barrett y por un antiguo estudiante de doctorado nuestro, Michael Schartner, hemos descubierto que la complejidad de la actividad cortical espontánea –medida por EEG– decae sistemáticamente tanto durante el sueño temprano como durante la anestesia.[14] También hemos hallado que la complejidad durante el sueño REM (o MOR, de movimientos oculares rápidos) es más o menos la misma que durante la vigilia consciente normal, algo bastante lógico, dado que, mientras estamos dormidos, es durante la fase REM cuando más probable resulta que tengamos sueños… y los sueños son conscientes. Massimini y su equipo habían hallado el mismo patrón de resultados con su indicador del ICP, lo que confirma aún más la idea de que estos indicadores monitorizan

realmente el nivel consciente más que el de vigilia.[15]

La capacidad de medir el nivel consciente de forma independiente del de vigilia no solo es importante desde el punto de vista científico, sino que es un avance potencialmente revolucionario para los neurólogos y para sus pacientes. El estudio de Massimini de 2013 ya mostraba que el ICP puede distinguir entre el estado vegetativo y el de mínima conciencia. Indicadores como el ICP pueden mostrarse tan potentes en ese contexto porque no se basan en la conducta externamente visible. La vigilia –o activación psicológica– se define en términos del comportamiento. En clínica, los neurólogos infieren típicamente el estado de vigilia a partir del hecho de que una persona reaccione a una estimulación sensorial, como un ruido fuerte o un pellizco en el brazo. Pero la conciencia se define en términos de la experiencia subjetiva interior, por lo que solo puede relacionarse de manera indirecta con aquello que es observable desde el exterior.

Los métodos clínicos convencionales para determinar el estatus consciente de un paciente con lesión cerebral siguen dependiendo aún de la manifestación conductual.[16] Normalmente, los neurólogos valoran no solo si un paciente responde a la estimulación sensorial (que es el marcador de la activación fisiológica), sino también si es capaz de interactuar con su entorno, ya sea respondiendo a órdenes o realizando comportamientos voluntarios. Cuando un paciente puede obedecer una solicitud en dos partes, y manifestar claramente cómo se llama y qué día es, inferimos que está plenamente consciente. El problema con este método es que algunos pacientes pueden poseer aún una vida interior que sean incapaces de expresar externamente. Las inferencias basadas puramente en el comportamiento pasan por alto esos casos y hacen que se diagnostique una ausencia de conciencia cuando, en realidad, esta sigue estando ahí.

Un ejemplo extremo de esto es el «síndrome de enclaustramiento», en el que la conciencia está plenamente presente pese a una completa parálisis del cuerpo. Esta rara afección puede ser consecuencia de un

daño en el tronco del encéfalo (o tallo cerebral), una región situada en la base del cerebro –y en el extremo superior de la médula espinal– que, entre otras funciones, interviene en el control de los músculos del cuerpo y de la cara.[17] Gracias a una peculiaridad anatómica, los pacientes con enclaustramiento pueden conservar todavía la facultad de efectuar movimientos oculares limitados, lo que abre un estrecho (y fácilmente inadvertido) canal conductual tanto para la diagnosis como para la comunicación. Un antiguo director de la revista *Elle,* Jean-Dominique Bauby, a quien se diagnosticó síndrome de enclaustramiento en 1995 a raíz de una hemorragia cerebral, escribió un libro entero de ese modo, *La escafandra y la mariposa.*[18] Pero los pacientes con enclaustramiento «completo» ni siquiera disponen de esos canales comunicativos, lo que hace que su diagnóstico sea más difícil todavía. Cuando lo conductual es la única guía, resulta demasiado fácil confundir un síndrome de enclaustramiento con una ausencia completa y permanente de la conciencia. Pero si introducimos a alguien como Bauby en un escáner cerebral, nos será fácil ver que su actividad cerebral general es normal casi por completo. En el estudio de Massimini de 2013, los pacientes con enclaustramiento presentaron valores de ICP indistinguibles de los individuos de control sanos de su misma edad, lo que indica que conservaban una conciencia totalmente intacta.

Los casos más dificultosos surgen en aquella zona gris entre la vida y la muerte, en situaciones como el estado vegetativo y el de mínima conciencia. En esas áreas liminares, las señales conductuales de conciencia pueden estar ausentes o parecer incoherentes, y el daño en el cerebro puede estar tan extendido que los escáneres craneales tampoco resulten concluyentes. Es ahí donde indicadores como el ICP podrían tener un efecto verdaderamente transformador. Cuando un paciente presenta un valor de ICP que indica la presencia de conciencia pese a que todo lo demás sugiera lo contrario, estamos ante alguien que merece ser reexaminado.

Marcello Massimini me contó hace poco un caso en el que la medición del ICP cambió por completo la situación. Un joven con heridas graves en la cabeza fue ingresado en un hospital de Milán. No respondía a preguntas ni instrucciones simples, lo que sugería una diagnosis de estado vegetativo. Pero su ICP era igual de elevado que el de una persona sana y plenamente consciente, una observación desconcertante, sobre todo, porque el suyo no era un caso de enclaustramiento. Tras una labor de rastreo, el equipo clínico logró localizar a un pariente, un tío que viajó a Italia desde el norte de África, donde aún vivía la familia del joven. Cuando ese tío suyo comenzó a hablar con su sobrino en árabe, se produjo una respuesta inmediata: sonreía con los chistes o incluso levantaba un pulgar en señal de aprobación cuando estaba viendo una película. Había estado consciente todo el tiempo; simplemente, no reaccionaba *al italiano*. El porqué es difícil de saber. Massimini cree que tal vez se tratara de un raro caso de «olvido cultural», como si el mundo italiano simplemente hubiera dejado de tener importancia alguna para él.[19] Fuera como fuere, la historia de este joven podría haber tenido un final muy distinto sin los reveladores ecos eléctricos medidos por el ICP.

La diagnosis de la conciencia residual en pacientes con lesión cerebral es un área de la medicina en rápido desarrollo. Además del ICP de Massimini, otros métodos están migrando actualmente desde los laboratorios hacia la práctica clínica. Mi favorito es uno basado en el famoso (para los neurólogos) experimento de la «casa del tenis», llevado a cabo por el neurocientífico Adrian Owen y su equipo en 2006. En dicho experimento, una mujer de veintitrés años que, tras un accidente de tráfico, había dejado de mostrar respuesta conductual alguna fue introducida en un escáner de IRMf, donde se le fueron dando varias instrucciones verbales. A veces, se le pedía que intentara imaginarse a sí misma jugando al tenis, mientras que, otras veces, se la animaba a que se imaginase andando por las habitaciones de su propia casa. A simple vista, el experimento puede parecernos bastante peculiar, porque los pacientes de ese tipo no responden a nada, ni, menos aún, a

instrucciones verbales complejas. Sin embargo, algunos estudios con personas sanas han mostrado que las regiones cerebrales que se activan al imaginar movimientos fluidos (como jugar al tenis) son muy distintas de aquellas otras que se estimulan cuando la persona se imagina desplazándose por espacios.* Lo asombroso del caso fue que la paciente de Owen exhibió exactamente ese mismo patrón de respuestas cerebrales, lo que indicaba que también ella estaba siguiendo activamente las instrucciones y reaccionando a ellas con imágenes mentales muy específicas. Es casi inconcebible que alguien pudiera hacer algo así estando inconsciente, por lo que Owen concluyó que el diagnóstico conductual de estado vegetativo que se le había hecho estaba equivocado, y que la joven se encontraba consciente, en realidad.[20] En la práctica, Owen y su equipo habían reconvertido un escáner cerebral en un instrumento con el que su paciente pudo interactuar con su entorno valiéndose de su cerebro en vez de su cuerpo.

Otros estudios posteriores han ido más allá incluso y han utilizado el método de Owen no solo para la diagnosis, sino también para la comunicación. En un estudio de 2010 dirigido por Martin Monti, un paciente ingresado en un hospital con un diagnóstico de estado vegetativo fue capaz de responder a preguntas de «sí o no» mediante el truco de imaginarse que estaba jugando al tenis para decir «sí», e imaginarse que estaba caminando por su casa para decir «no». Un laborioso mecanismo de comunicación, sin duda, pero un avance revolucionario para quienes no tienen otro modo de que se les entienda.[21]

¿Cuántas personas conscientes, aunque no muestren respuesta fisiológica, languidecerán olvidadas en salas hospitalarias de neurología y en geriátricos? Difícil saberlo. El método de Owen, más antiguo que el ICP, también ha sido investigado con mayor frecuencia, y un análisis reciente sugiere que entre un 10 % y un 20 % de los pacientes en estado vegetativo podrían conservar algún tipo de conciencia encubierta, lo que representaría muchos miles de casos en todo el mundo.[22] E incluso eso

podría ser una subestimación. Para pasar el test de Owen, los pacientes tienen que entender el idioma en el que les están hablando y mantener períodos prolongados de producción de imágenes mentales, algo que algunos de ellos podrían no ser ya capaces de hacer, a pesar de seguir conscientes. Ahí es donde nuevos métodos como el del ICP tienen especial significación, ya que prometen brindarnos la capacidad de detectar una consciencia residual sin obligar al paciente a hacer nada en absoluto. Justo lo que un verdadero medidor de conciencia debe hacer.

El concepto de «nivel de conciencia», tal como lo he usado hasta el momento, identifica cambios relativamente globales en cuanto a lo consciente que está un individuo, por ejemplo, entre su vida despierta normal y un estado vegetativo o uno de anestesia general. No obstante, existen otras maneras de concebir el nivel de conciencia. ¿Es menos consciente un bebé que un adulto? ¿Es menos consciente una tortuga que un humano de cualquier edad?

Existe, claro está, un riesgo en pensar de ese modo. Esta clase de preguntas nos invitan a presuponer que cualquier forma de conciencia que diverge de la del humano adulto sano es menor o inferior en algún sentido. Esta mentalidad es sintomática del excepcionalismo humano que tan repetidas veces ha lacrado a la biología, al tiempo que oscurecía la historia del pensamiento en todos los terrenos y culturas. La conciencia tiene muchas propiedades y es un error confundir el conjunto particular de dichas propiedades característico de las personas adultas sanas con la naturaleza esencial de la conciencia en todas sus formas, presuponiendo que la humana ocupa la cima de una sola escala unidimensional. No cabe duda de que las experiencias conscientes surgen con el tiempo, ya sea durante el desarrollo de un animal (humano o no), ya sea a lo largo de la inmensidad temporal de la evolución. Pero de ahí a describir cualquiera de esos dos procesos como si siguiera una única línea de desarrollo, o como si su punto culminante fuese el ideal humano adulto de modo de ser tú (o yo), va un muy largo trecho. He ahí un posible límite a la analogía entre la conciencia y la

temperatura con la que empecé este capítulo.[23]

Una cuestión relacionada con la anterior es la de si la conciencia es un «todo o nada» (la luz está encendida o no lo está) o si se manifiesta por «grados» (sin que haya una línea perfectamente clara entre la presencia y la ausencia de conciencia). Esta es una pregunta que sirve tanto para la aparición de la conciencia en la evolución general o en el desarrollo individual, como para cuando volvemos de la inconsciencia en la que nos encontrábamos durante la anestesia general o mientras dormíamos sin soñar. Pero pese a que se trata de un interrogante muy llamativo, creo que está mal planteado. La distinción entre una conciencia binaria (de «todo o nada») y otra de «grados» no tiene por qué ser absoluta, de una cosa o la otra. Ya sea en la evolución, en el desarrollo individual, en la vida cotidiana o en la sala de neurología de un hospital, siempre prefiero pensar en términos de transiciones más o menos rápidas desde la ausencia total de conciencia a la presencia de, como mínimo, cierta experiencia consciente, la cual se manifiesta entonces en diferentes grados, tal vez incluso en diferentes dimensiones, desde el momento en que las luces interiores comienzan, cuando menos, a emitir algún tenue destello.

Pongamos el ejemplo de una persona adulta. ¿Su nivel de conciencia es mayor (o menor) cuando está soñando que cuando está sentada a su mesa de trabajo tras un almuerzo pesado, en un estado de sopor? No hay respuestas directas para preguntas como esta. Soñar puede ser un ejercicio «más consciente» en ciertos sentidos (por ejemplo, en cuanto a la intensidad de su fenomenología perceptiva), pero «menos consciente» en otros (por ejemplo, en cuanto al grado de comprensión reflexiva de lo que está ocurriendo).[*]

Una consecuencia importante de tomarse en serio la multidimensionalidad de los niveles de conciencia es que, de ese modo, desaparecen las distinciones drásticas entre el *nivel* de la conciencia y el *contenido* de esta. Pierde su sentido el diferenciar radicalmente entre cuán consciente estás y de qué eres consciente. Es posible, pues, que con

un solo indicador universal «de talla única» de la conciencia, como el que esperaríamos obtener si nos tomáramos la analogía de la temperatura demasiado al pie de la letra, no tengamos nunca suficiente.

Un ejemplo de la interacción entre el nivel y el contenido conscientes nos lo proporciona un estudio sobre la actividad cerebral en el estado psicodélico que llevamos a cabo hace unos años. Entre sus muchos usos, las drogas o fármacos psicodélicos brindan oportunidades únicas para la ciencia de la conciencia porque inducen alteraciones profundas de los contenidos conscientes a partir de una simple intervención farmacológica en el cerebro.[24]

El químico suizo Albert Hofmann, inventor de la dietamida de ácido lisérgico (el LSD), escribió un relato de su viaje de vuelta a casa desde los laboratorios de la empresa farmacéutica Sandoz, en Basilea, el 19 de abril de 1943, que nos da una buena idea de lo extraordinarias que pueden ser dichas alteraciones. Ese día, que hoy se recuerda con el nombre de «Día de la Bicicleta», Hofmann decidió ingerir una pequeña cantidad de la sustancia que acababa de descubrir. Al poco tiempo, comenzó a sentirse muy raro, así que optó por marcharse a casa en bicicleta. Tras conseguir llegar a su destino, no se sabe muy bien cómo, mientras lo asaltaban toda clase de experiencias perturbadoras, y tras creer que estaba perdiendo el juicio, se tumbó en el sofá y cerró los ojos.

> […] Poco a poco, empecé a disfrutar de unos colores y unos juegos de formas sin precedentes que no se me iban de los ojos cerrados. Me invadía una oleada de imágenes caleidoscópicas y fantásticas que se iban alternando y diversificando, abriéndose y cerrándose en círculos y espirales, explotando en fuentes de colores, reorganizándose e hibridándose en un flujo constante […].[25]

En el estado psicodélico, unas alucinaciones perceptivas muy vivas vienen a menudo acompañadas de inusuales experiencias de la yoidad –descritas por quienes las experimentan como una «disolución del ego»– en las que los límites entre el yo y el mundo (y otras personas) parecen

cambiar o disiparse. Estas variaciones con respecto a la experiencia consciente «normal» son tan intensas que el estado psicodélico podría representar un cambio no ya en los *contenidos* conscientes, sino también en el *nivel* consciente en general. Esa, al menos, fue la idea que nos propusimos testar en colaboración con Robin Carhart-Harris, del Imperial College de Londres, y Suresh Muthukumaraswamy, de la Universidad de Auckland.

En abril de 2016, Robin y yo participamos en un congreso celebrado al pie de las montañas de Santa Catalina, en las afueras de Tucson (Arizona). Ambos habíamos sido invitados para presentar ponencias sobre nuestras investigaciones y estábamos aprovechando la ocasión para explorar de qué modo el interés del uno y el otro por la conciencia podría encontrar puntos de coincidencia en el contexto de la psicodelia. Por entonces, acababan de reanudarse las investigaciones científicas y médicas sobre el LSD y sobre otros compuestos psicodélicos como la psilocibina (el ingrediente activo de las setas alucinógenas) tras décadas de abandono. Tras la autoexperimentación de Hofmann, había habido un breve florecimiento de los estudios dedicados a explorar el potencial del LSD para el tratamiento de una serie de trastornos psicológicos, incluidos la adicción y el alcoholismo, con resultados muy prometedores. Pero la subsiguiente adopción del LSD como droga recreativa y como símbolo de rebelión –promocionada por Timothy Leary, entre otros– hizo que buena parte de esas investigaciones se clausuraran a lo largo de la década de 1960. Hubo que esperar a la década de 2000 para ver el reinicio de algún estudio sustancial al respecto; mientras tanto, se había perdido una generación entera de avances científicos.[26]

En el nivel neuroquímico, las sustancias psicodélicas clásicas –el LSD, la mescalina y la dimetiltriptamina (la DMT, ingrediente activo del brebaje alucinógeno sudamericano conocido como ayahuasca)– actúan fundamentalmente afectando al sistema serotonínico del cerebro. La serotonina es uno de los principales neurotransmisores cerebrales

(sustancias químicas que inundan los circuitos del cerebro e influyen en cómo se intercomunican las neuronas). Las drogas psicodélicas influyen en el sistema serotonínico porque se vinculan con fuerza a un receptor de serotonina concreto, el 5-HT_{2A}, que se encuentra distribuido por buena parte del cerebro. Uno de los principales desafíos a los que se enfrenta la investigación sobre sustancias psicodélicas es el de entender cómo logran estas intervenciones farmacológicas de bajo nivel modificar los patrones globales de la actividad cerebral hasta el punto de producir profundos cambios en la experiencia consciente.

El equipo de Robin ya había descubierto que el estado psicodélico implica unas llamativas alteraciones de la dinámica cerebral en comparación con una situación «placebo» de control.[27] Las redes de regiones del cerebro que normalmente se coordinan durante la actividad de este –llamadas «redes en estado de reposo»– se desacoplan en ese estado, mientras que otras regiones que habitualmente funcionan de un modo más o menos independiente se enlazan entre sí. La imagen de conjunto resultante es de una especie de ruptura de los patrones de conectividad por los que se caracteriza el cerebro en condiciones normales. Robin concluyó que esas rupturas podrían explicar ciertos rasgos distintivos del estado psicodélico, como la disolución de los límites entre el yo y el mundo, o las experiencias sinestésicas.

Robin y yo nos dimos cuenta de que los datos que él había estado recopilando eran ideales para los análisis de complejidad algorítmica que, en mi equipo de Sussex, habíamos estado aplicando al estado dormido y a la anestesia. En particular, algunos de sus escáneres cerebrales se habían realizado con magnetoencefalografía (MEG), que proporciona la clase de alta resolución temporal y cobertura global del cerebro que nosotros necesitábamos.[28] Ellos habían empleado la MEG para medir la actividad cerebral en personas voluntarias que habían ingerido psilocibina, LSD o dosis bajas de ketamina. (Las dosis altas de ketamina actúan como un anestésico, pero las bajas tienen un efecto más bien alucinógeno). Nosotros podíamos usar esos datos para

responder a la pregunta siguiente: ¿qué les ocurre a las mediciones del nivel consciente cuando los contenidos conscientes cambian tan drásticamente como lo hacen durante un «viaje» psicodélico?

De vuelta en Sussex, Michael Schartner y Adam Barrett calcularon los cambios en la complejidad algorítmica de la señal de la MEG entre los tres estados psicodélicos en muchas regiones diferentes del cerebro. Los resultados fueron tan claros como sorprendentes: la psilocibina, el LSD y la ketamina inducen *incrementos* en comparación con un placebo de control.[29] Aquella fue la primera vez que alguien observaba un aumento en un indicador del nivel consciente con relación a una situación base de reposo despierto. Todas las comparaciones previas, ya fueran con el sueño, la anestesia o los trastornos de la conciencia, habían arrojado disminuciones en esos indicadores.

Para entender lo que significa este resultado, recuerda que los indicadores de complejidad algorítmica que utilizamos vienen a ser unas medidas de la aleatoriedad o la «diversidad de las señales» de un cerebro. Una secuencia completamente aleatoria tendrá la máxima complejidad algorítmica, la mayor diversidad posible. Así pues, nuestros resultados complementaban los estudios previos de Robin porque mostraban que la actividad cerebral en el estado psicodélico se vuelve más aleatoria con el tiempo, lo que concuerda con la reorganización espontánea de la experiencia perceptiva que las personas suelen referir cuando están en pleno «viaje».[30] También arrojan luz sobre cómo el nivel y el contenido conscientes están relacionados entre sí. He ahí un ejemplo de un indicador del *nivel* consciente reaccionando a las variaciones generalizadas de los *contenidos* de la conciencia características del estado psicodélico. El hecho de que una medida del nivel consciente sea también sensible a los cambios en el contenido consciente pone de manifiesto que no son aspectos de la conciencia independientes.

De los resultados de nuestros análisis psicodélicos se desprendía también una perspectiva preocupante. ¿Conduciría una actividad

cerebral aleatoria máxima –según se mide con la complejidad algorítmica– a una experiencia psicodélica máxima? ¿O llevaría, más bien, a un «nivel» de conciencia diferente, de algún otro tipo? La extrapolación parece improbable. Más bien se diría que un cerebro en el que todas las neuronas se disparasen caóticamente no daría origen a experiencia consciente alguna, igual que el *free jazz* puede, en algunos momentos, dejar de ser música.

En este caso, la cuestión es que la complejidad algorítmica es una mala aproximación a lo que se suele entender por complejidad cuando se dice que algo «es complejo». Intuitivamente, complejidad y aleatoriedad no son la misma cosa. Una noción más satisfactoria de la primera sería aquella que la sitúa en un terreno intermedio entre el orden y el desorden, pero no en el punto extremo de este último. Es Nina Simone y Thelonius Monk, no la Bonzo Dog Doo-Dah Band.[*] ¿Qué sucede si, en vez de con esa otra, nos quedamos con esta forma más sofisticada de concebir la complejidad?

Precisamente eso hicieron Giulio Tononi y mi antiguo jefe y mentor, Gerald Edelman, en un artículo publicado en 1998 en la revista *Science*.[31] Todavía recuerdo cuando lo leí por primera vez, hace ahora una veintena de años. Aquello marcó un hito en mi manera de pensar sobre la conciencia y fue, en buena medida, lo que me llevó a trabajar en el Instituto de Neurociencias de San Diego.

En vez de centrarse –como hace el enfoque de los «correlatos neurales de la conciencia» (CNC)– en una única experiencia consciente a modo de ejemplo (como la experiencia de «ver el color rojo»), Tononi y Edelman se preguntaron qué es lo característico de las experiencias conscientes en general. E hicieron una apreciación muy simple, pero también muy profunda, al respecto: que las experiencias conscientes –*todas* las experiencias conscientes– son tanto *informativas* como *integradas*. Desde ese punto de partida, plantearon hipótesis sobre la base neural de *todas* las experiencias conscientes, y no solo de experiencias concretas, como ver el rojo, sentir celos o sufrir un dolor de muelas.

La idea de la conciencia como un fenómeno informativo a la vez que integrado requiere de cierto desarrollo conceptual.

Empecemos por la información. ¿Qué significa que las experiencias conscientes sean «informativas»? Edelman y Tononi no querían decir con esto que lo fueran en el mismo sentido en que un periódico puede ser informativo, sino para indicar que, aunque puedan parecer triviales en un principio, subyace a ellas una gran riqueza. Las experiencias conscientes son informativas porque cada una de ellas es diferente de todas las demás que jamás hayas tenido, tendrás o pudieras tener.

Si miro más allá de la mesa de trabajo que tengo delante y a través de la ventana, me doy cuenta de que nunca antes he experimentado *esta* configuración exacta de tazas de café, pantallas de ordenador y nubes, una experiencia que es aún más diferenciada, si cabe, cuando la sumamos a todas las demás percepciones, emociones y pensamientos que están simultáneamente presentes de fondo en mi universo interior. En cualquier instante dado, tengo precisamente *una* experiencia consciente de la *inmensidad* de experiencias conscientes *posibles*. Cada experiencia consciente, pues, produce una ingente reducción de la incertidumbre, ya que es *esa* experiencia la que se está teniendo y no *aquella,* ni *aquella* otra, etcétera. Y, en matemáticas, ese justamente es el significado de información: reducción de incertidumbre.

Lo informativa que una experiencia consciente particular sea no está en función de lo rica o detallada que es, ni de cuánto ilustra a la persona que la está teniendo. Escuchar a Nina Simone y comer fresas al mismo tiempo mientras uno va subido en una montaña rusa descarta tantas experiencias alternativas potenciales como el estar sentado con los ojos cerrados en una habitación en silencio, sin experimentar casi nada. Cada experiencia reduce exactamente la misma cantidad de incertidumbre con respecto al abanico de experiencias posibles.

Desde ese punto de vista, el «modo de ser» de cualquier experiencia consciente no se define tanto por lo que *es* como por todas las cosas posibles, aunque inmaterializadas, que *no* es. Una experiencia de visión

de un rojo puro es como es, no por una propiedad intrínseca de «rojez», sino porque el rojo no es el azul, ni el verde, ni ningún otro color; ni olor, ni pensamiento, ni sentimiento de arrepentimiento, ni, para el caso, ninguna otra forma de contenido mental. La rojez es rojez por todas esas otras cosas que no es, y lo mismo puede decirse de todas las demás experiencias conscientes.

Ahora bien, puntuar alto en el apartado de la información por sí solo no basta. Las experiencias conscientes no son únicamente muy informativas, sino que también están *integradas.* El significado exacto de que la conciencia esté «integrada» sigue siendo habitual objeto de debate. El caso es que no experimentamos los colores de manera separada de sus formas, ni los objetos de un modo independiente de sus fondos. Los múltiples elementos distintos que componen mi experiencia consciente en este preciso instante –ordenadores y tazas de café, pero también el ruido de una puerta que se cierra en el pasillo, y mis pensamientos sobre qué escribir a continuación– parecen entrelazados de un modo ineludible y fundamental como si fueran aspectos de una sola escena consciente que los abarca a todos.[32]

La propuesta maestra de Tononi y Edelman fue postular que, si toda experiencia consciente es informativa a la vez que unificada a un nivel fenomenológico, *entonces cabe suponer que los mecanismos neurales que subyacen a las experiencias conscientes también exhiban esas dos propiedades,* es decir, que por el hecho mismo de que manifiestan ambas propiedades, los mecanismos neurales no solo se correlacionen con los rasgos fenomenológicos de toda experiencia consciente, sino que, de hecho, los expliquen.

¿Qué significa que un mecanismo sea integrado e informativo al mismo tiempo? Abstraigámonos por un momento del caso del cerebro y pensemos en un sistema que consista en un gran número de elementos interactivos entre sí, sin que nos importe cuáles sean tales elementos. Como se muestra en la ilustración de la página siguiente, para cualquier sistema de ese tipo, podemos definir una escala con dos puntos

extremos. En uno de ellos (a la izquierda), todos los elementos se comportan aleatoria e independientemente, como las moléculas de un gas. Esta clase de sistema posee información máxima –máxima aleatoriedad– pero nula integración, porque cada elemento es independiente de todos los demás.

En el otro extremo (a la derecha), todos los elementos hacen exactamente lo mismo, por lo que el estado de cada elemento está complemente determinado por el estado de los demás del sistema. Aleatoriedad cero. Sería como la disposición de los átomos en una estructura cristalina, en la que la posición de cualquier átomo individual está totalmente determinada por la estructura misma, que se define por las posiciones de todos los demás átomos. Esta clase de disposición posee una integración máxima, pero casi ninguna información, porque son muy pocos los estados posibles de ese sistema.

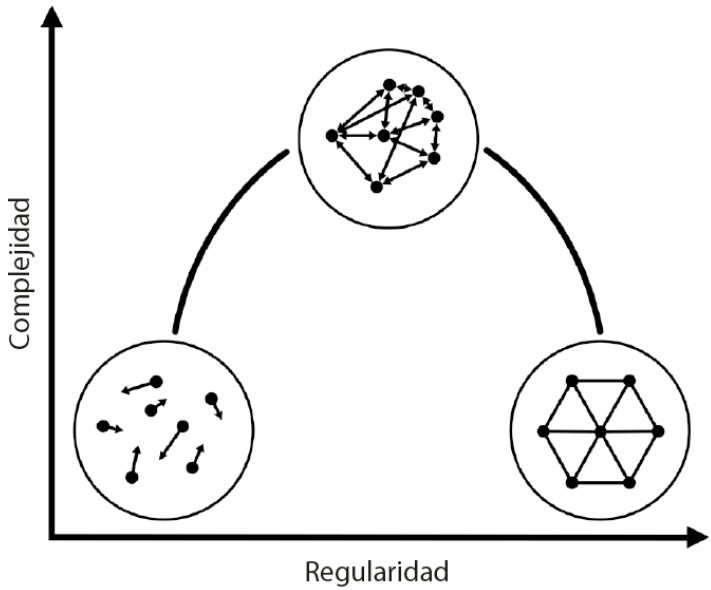

Fig. 2. La relación entre complejidad y regularidad

Entremedias se sitúan aquellos sistemas en los que los elementos individuales pueden hacer cosas diferentes, pero en los que existe cierto

grado de coordinación que hace que el sistema se comporte hasta cierto punto «como un todo». Ese es el terreno en el que encontramos tanto integración como información. También es la zona intermedia entre el orden y el desorden, y es donde solemos calificar los sistemas de «complejos».

Esas descripciones arrojan luz sobre la base neural de la conciencia cuando las aplicamos al caso del cerebro.

En un cerebro que estuviera al máximo de riqueza informativa, todas las neuronas se comportarían independientemente y se activarían al azar, como si estuvieran completamente desconectadas de las otras. En un cerebro así, los indicadores de la complejidad algorítmica (como el de la complejidad LZW, por ejemplo) arrojarían puntuaciones muy altas. Pero ese cerebro –con mucha información, pero nula integración– no sustentaría estado consciente alguno. En el extremo opuesto, en un cerebro ordenado al máximo, todas las neuronas estarían haciendo exactamente lo mismo, tal vez activándose juntas al unísono, algo parecido a lo que ocurre durante los ataques epilépticos globales. La complejidad algorítmica en ese caso sería muy baja. Ese cerebro también carecería de conciencia, pero por distinto motivo: mucha integración, nula información.

Así pues, un indicador idóneo del nivel consciente sería aquel que registrara no la información en sí, sino más bien cómo la información y la integración se expresan de forma conjunta. Un indicador así –una medida de la complejidad en el verdadero sentido– ejemplificaría la manera de abordar la conciencia desde el enfoque del problema real al entrelazar explícitamente propiedades del mecanismo con propiedades de la experiencia.

Como ya hemos visto, las aproximaciones a la complejidad algorítmica –la complejidad LZW, por ejemplo– no sirven bien a ese propósito. Nos dicen mucho sobre la información, pero nada sobre la integración. El ICP lo hace un poco mejor. Para que la puntuación en la escala del ICP sea alta, el pulso de energía inyectado mediante la EMT

tiene que generar un patrón de actividad cerebral que sea difícil de comprimir, lo que indicará un elevado nivel de información. El pulso también tiene que desplazarse muy a lo largo y ancho del córtex para generar un «eco» del que luego se pueda evaluar la compresibilidad. No obstante, aunque esta difusión cortical sugiere integración, seguirá quedándose corta respecto a lo que idealmente exigiríamos de un indicador así. La medida del ICP depende de que la actividad cerebral esté integrada de una forma bastante imprecisa –pues, de otro modo, no habría eco–, pero no mide la integración con la misma precisión cuantitativa con la que sí mide la información. Lo que buscamos son indicadores que sean directamente sensibles tanto a la integración como a la información a partir de unos mismos datos, de un mismo modo y a un mismo tiempo.

Hay varias medidas que cumplen esos criterios (en teoría, al menos). Allá por la década de los noventa, Tononi y Edelman, junto con su colega Olaf Sporns, idearon un indicador que denominaron «complejidad neuronal»[33] y, diez años más tarde, yo derivé el mío propio, al que llamé «densidad causal»,[34] y lo hicimos valiéndonos de un tipo distinto de cálculos matemáticos. Partiendo de esas bases, una serie de indicadores más nuevos –algunos de los cuales conoceremos en el próximo capítulo– han ido tomando desarrollos cada vez más sofisticados. Todas estas medidas tratan de cuantificar, de un modo u otro, qué punto ocupa un sistema en el terreno intermedio entre el orden y el desorden, que es donde podemos encontrar la información integrada. El problema, sin embargo, es que ninguno ha funcionado demasiado bien cuando se ha aplicado a los datos de imágenes cerebrales reales.

La situación actual resulta curiosa en cierto sentido. Cabría esperar que los indicadores más fieles a los principios teóricos funcionaran mejor en la práctica que aquellos otros, como la complejidad algorítmica, que solo se vinculan de forma indirecta con la teoría subyacente. Pero no es eso lo que vemos en realidad, así que ¿qué está

pasando? Una posibilidad es que la teoría en sí esté equivocada. Yo, sin embargo, tengo la intuición de que lo único que necesitamos es perfeccionar la parte matemática para que los indicadores hagan lo que nos gustaría que hicieran, además de desarrollar técnicas de imagen cerebral mejoradas que produzcan el tipo adecuado de datos para tales indicadores.[35]

Todavía prosigue, pues, la búsqueda de un auténtico medidor de la conciencia. Conviene recalcar que los avances realizados hasta la fecha han sido considerables. Hoy está ampliamente reconocido el hecho de que el nivel de conciencia no es lo mismo que el nivel de vigilia, y disponemos ya de una serie de indicadores cerebrales del nivel consciente que funcionan impresionantemente bien a la hora de monitorizar diferentes estados globales de conciencia y detectar la consciencia residual en pacientes con lesiones cerebrales. El ICP de Massimini ha tenido especial importancia en ese sentido. Es clínicamente útil y, además, está firmemente asentado sobre sólidos principios teóricos de lo que es la información y la integración, por lo que conecta en la práctica los mecanismos neuronales con las propiedades universales de la experiencia consciente, muy al estilo del enfoque del «problema real». Continuamente surgen, además, otros indicadores basados en principios diferentes aunque relacionados con los anteriores. Por otra parte, ciertas aproximaciones fáciles de usar, como la estimación de la complejidad algorítmica de datos cerebrales espontáneos, nos están revelando fascinantes conexiones entre el nivel y el contenido conscientes.[36]

Pero sigue en pie una pregunta fundamental: ¿se parece la conciencia más a la *temperatura,* es decir, a algo que se puede reducir a, e identificar con, una propiedad básica del universo físico (o informacional)? ¿O es más bien como la *vida,* una constelación de múltiples propiedades diferentes, cada una con su propia explicación en términos de mecanismos subyacentes? Los enfoques para la medición de la conciencia que hemos visto hasta el momento se inspiran en esa

analogía de la temperatura, pero intuyo que, al final, pueden encajar mejor en la de la vida. Para mí, la «integración» y la «información» son propiedades generales de la mayoría –si no la totalidad– de las experiencias conscientes. Pero eso no quiere decir que la conciencia *es* información integrada igual que sí podemos afirmar que la temperatura *es* la energía cinética molecular media.

Para ver qué me induce a tener tal intuición, hará falta que estiremos la analogía entre la conciencia y la temperatura hasta donde esta pueda dar de sí, para ver si llega a romperse y cuándo lo hace. Es hora de que conozcamos la «teoría de la información integrada» referida a la conciencia.

3. FI

Es julio de 2006 y estoy en Las Vegas, comiéndome un helado italiano con Giulio Tononi. Nos encontramos en el hotel Venetian y la verdad es que no me entero prácticamente de nada de lo que está pasando. He llegado en avión el día antes desde Londres, y allí, en las profundidades del Venetian, siempre está uno como en un anochecer perpetuo, con sus estrellas falsas que empiezan a relucir tenues sobre un cielo cerúleo no menos simulado, y sus góndolas de pega deslizándose entre aquellos *palazzi* de mentira también. Lo mantienen así para que la gente se quede allí gastándose el dinero, en una especie de eterna hora de la *apericena,* habiendo perdido la noción del tiempo transcurrido. Tengo *jet lag,* estoy un poco achispado por la bebida –antes de pasar a esa zona habíamos tenido una larga sobremesa– y llevamos ya horas discutiendo los detalles de la ambiciosísima «teoría de la información integrada» (TII) referida a la conciencia. Es una idea de Tononi que aborda directamente, como ninguna otra teoría de raigambre neurocientífica, el problema difícil de la conciencia. La TII dice que la experiencia subjetiva es una propiedad de unos patrones de causa y efecto, que la información es tan real como la masa o la energía y que incluso los átomos pueden ser un poquito conscientes.[1]

No se trata de un debate igualado, ni de lejos. Yo he dedicado buena parte de mi tiempo esa tarde a defender un artículo mío reciente en el que critico una versión anterior de su teoría.[2] Giulio está esforzándose –de manera delicada pero persistente– por explicarme por qué estoy equivocado. No estoy seguro de si es por el *jet lag,* por el vino o por la

implacable lógica de Giulio, pero lo cierto es que no estoy tan convencido de mis posiciones como lo estaba mientras volaba hacia allí. A la mañana siguiente, decido que lo pensaré más a fondo, que trataré de comprenderlo mejor, que me lo prepararé más… y que beberé menos.

La TII me pareció fascinante entonces y me lo sigue pareciendo ahora, porque ejemplifica muy bien la analogía entre la conciencia y la temperatura. Según la TII, la conciencia simplemente *es* información integrada. En sus planteamientos, la teoría da completamente la vuelta a ciertas intuiciones muy arraigadas sobre la relación entre mente y materia, y sobre cómo se entrelaza la conciencia con el tejido del universo en general.

En aquel 2006, la TII todavía no era muy conocida. Ahora, es una de las teorías más destacadas, pero también una de las más controvertidas, de la ciencia de la conciencia. Ha sido objeto del elogio de algunos de los más grandes nombres de este campo, aparte de Tononi. Christof Koch, antiguo adalid del enfoque de los CNC, la calificó de «paso de gigante hacia la resolución definitiva del antiguo problema mente-cuerpo».[3] Pero por su carácter ambicioso y su notoriedad, ha chocado también con una considerable resistencia. Una de las razones de esa reacción adversa radica en que se trata de un enfoque profundamente matemático y desinhibidamente complejo. Algo que, por supuesto, no tiene por qué ser malo: nadie dijo que descifrar el enigma de la conciencia fuera a ser sencillo. Otra crítica que se le hace es que las tesis que propone son tan contraintuitivas que la teoría debe de estar equivocada. Pero esta también es una intuición muy peligrosa en la que basarse cuando se trata de encarar un fenómeno tan desconcertante como la conciencia.[4]

Para mí, el principal problema es que los extraordinarios postulados de la TII requieren de unas extraordinarias evidencias empíricas, pero precisamente lo ambicioso de la propia TII –su pretensión de resolver el problema difícil– hace que en la práctica sea imposible contrastar sus hipótesis más características. Las extraordinarias pruebas que se necesita

reunir son imposibles de obtener. Por suerte, no todo está perdido. Como explicaré, algunas predicciones de la TII sí se pueden testar, al menos en principio. Y existen interpretaciones alternativas de la propia teoría, más alineadas con el enfoque del problema real que con el del problema difícil, que están impulsando el desarrollo de nuevos indicadores del nivel consciente aplicables en la práctica y, al mismo tiempo, fundamentados en principios teóricos.

Como su nombre indica, los conceptos de «información» e «integración» forman el núcleo de la TII. La teoría se fundamenta en ideas sobre la medición del nivel consciente que ya vimos en el capítulo anterior, pero lo hace de un modo realmente distintivo.

El elemento central de la TII es un único indicador llamado «Φ» (la letra griega «fi»). En su definición más sencilla, Φ vendría a ser una medida de lo mucho que un sistema es «más que la suma» de sus partes, en lo que a información respecta. ¿Cómo puede ser un sistema más que la suma de sus partes? Una bandada de aves nos proporciona una analogía aproximada: la bandada parece ser más que la suma de las aves que la componen, pues parece tener «vida propia». La TII adopta esa idea y la traslada al ámbito de la información. En la TII, Φ mide la cantidad de información que genera un sistema «como un todo» que exceda la cantidad sumada de información que generan sus partes de forma independiente. Ese es el pilar sobre el que se apoya la tesis principal de la teoría, que es que *un sistema es consciente en la medida en que su todo genera más información que sus partes.*[5]

Debe repararse en que no es una tesis sobre una correlación, ni es una propuesta (al estilo del enfoque del problema real) sobre cómo las propiedades mecánicas de un sistema explican las propiedades de su fenomenología. Es más bien una tesis sobre una identidad. Según la TII, el nivel de Φ es intrínseco de un sistema (es decir, que no depende de un observador externo) y es idéntico a la cantidad de conciencia asociada a ese sistema. ¿Que Φ es alto? Mucha conciencia. ¿Que Φ es cero? Nula conciencia. De ahí que la TII sea la máxima expresión de una

concepción de la conciencia análoga a la de la temperatura.[6]

¿Qué se necesita para que haya un Φ alto? Aunque la idea de base ya debería resultarnos conocida tras haber leído el capítulo previo, existen también ciertas diferencias importantes, así que merece la pena empezar por el principio.

Imagina una red de «neuronas» artificiales simplificadas, cada una de las cuales puede estar «encendida» o «apagada». Para que Φ sea alto, la red tiene que satisfacer dos condiciones principales. En primer lugar, el estado global de la red –la red «como un todo»– tiene que implicar el descarte de un gran número de estados globales alternativos posibles. Eso es *información,* y se corresponde con la observación de la fenomenología según la cual toda experiencia consciente descarta muchísimas otras experiencias conscientes alternativas posibles. En segundo lugar, debe haber más información cuando se considera el sistema como un todo que cuando, simplemente, se suman las partes en las que se subdivide (sus neuronas individuales o sus grupos de neuronas). Eso es *integración* y se corresponde con la observación según la cual todas y cada una de las experiencias conscientes forman conjuntos unificados, y se experimenta cada una «en bloque». Φ es una manera de asignarle a un sistema un número que mida cuánto puntúa en esas dos dimensiones a la vez.

Son muchos los motivos por los que un sistema puede no presentar un Φ alto.[7] Uno es que puntúe bajo en información. Un ejemplo mínimo es el de un único fotodiodo: un sensor simple de luz que puede estar «encendido» o «apagado». Este tiene Φ bajo o cero porque la información que proporciona su estado en cualquier momento dado es muy escasa. Sea cual sea el estado en el que esté (uno o cero, encendido o apagado), solo descarta como máximo una alternativa (cero o uno). Un fotodiodo solitario transmite a lo sumo un «bit» de información.[*]

Un sistema puede tener un Φ bajo si puntúa poco en integración. Imaginemos un conjunto numeroso de fotodiodos, como, por ejemplo, los del sensor de la cámara de nuestro móvil. El estado global del

sistema es el de todo el conjunto y eso puede transmitir una gran cantidad de información. Si el grupo de sensores es lo bastante grande, entrará en un estado global diferente por cada estado del mundo distinto con el que interactúe, que es la razón por la que las cámaras nos son útiles. Pero esa información global no importa para el sensor en sí. Los fotodiodos individuales del sensor son todos causalmente independientes los unos de los otros: su estado depende solamente del nivel de luz que cada uno de ellos detecte. Si le cortas al sensor un subgrupo de fotodiodos (causalmente independientes), este seguirá funcionando igual de bien.[8] La información transmitida por el conjunto de sensores como un todo no es mayor que la transmitida por todos sus sensores –todos los fotodiodos– funcionando de manera independiente. Eso significa que la información que genera no es mayor que la suma de sus partes y, por lo tanto, que su Φ será también cero.

Otro instructivo ejemplo de Φ cero es el de las situaciones denominadas de «cerebro dividido». Imagina una red dividida en dos mitades completamente separadas. Cada mitad puede tener un Φ distinto de cero, pero la red en su conjunto tendrá Φ cero. Esto es debido a que existe un modo de dividir la red en partes –las dos mitades– de tal manera que el todo no sea más que la suma de dichas partes. Este ejemplo pone de relieve cómo Φ depende del modo óptimo en que se puede «trocear» un sistema para minimizar la diferencia entre lo que hace el todo y lo que hacen las partes. Ese es uno de los aspectos distintivos de la TII que la diferencia de los indicadores de la complejidad descritos en el capítulo previo.

Este ejemplo implica también que los cerebros divididos *reales* –a raíz de un procedimiento quirúrgico que divide los hemisferios corticales, como el que se aplica a algunos casos de epilepsia que no se pueden tratar de otro modo–[9] pueden albergar dos «conciencias» independientes, pero no habrá ningún ente consciente unitario que abarque ambos hemisferios. Lo digo en el mismo sentido en que tú y yo somos ambos seres conscientes, pero no existe ningún ente colectivo

consciente que nos englobe a los dos, porque se nos puede dividir (informacionalmente hablando) al uno del otro justo por la mitad.

Siguiendo un poco más con los cerebros reales, digamos también que la TII explica a la perfección toda una serie de observaciones sobre el nivel consciente.[10] En el capítulo anterior, vimos que el cerebelo no parece participar mucho en la conciencia, pese a contener tres cuartas partes del total de neuronas que hay en el encéfalo. La TII explica que esto sea así, porque equipara la anatomía del cerebelo a la del conjunto de sensores de tu cámara de móvil: un número increíblemente grande de circuitos semiindependientes y pobres en Φ. La corteza cerebral, por el contrario, es toda ella un cableado de conductores densamente interconectados que tiende a asociarse a un Φ alto. Entonces, ¿por qué se desvanece la conciencia durante estados como el de dormir sin soñar, la anestesia o el coma, si ese cableado no varía? La TII nos dice que, en tales estados, la capacidad de las neuronas corticales para interactuar unas con otras remite de tal modo que Φ se desploma.

La TII representa un enfoque «axiomático» de la conciencia. Parte de principios teóricos, más que de datos experimentales. En lógica, los axiomas son enunciados verdaderos por sí mismos, en tanto en cuanto suscitan un consenso general y no requieren de justificación adicional. Un buen ejemplo es el siguiente, tomado del filósofo griego Euclides: «Dos formas que ocupan exactamente el mismo espacio son la misma forma». La TII postula axiomas sobre la conciencia –sobre todo, el de que las experiencias conscientes son tanto integradas como informativas– y usa esos axiomas para derivar tesis sobre las propiedades que cabe suponerles a los mecanismos subyacentes a esas experiencias.[11] Basándonos en la TII, *cualquier* mecanismo que presente esas propiedades, sea un cerebro o no, sea biológico o no, tendrá un Φ distinto de cero y, por lo tanto, poseerá conciencia.

Hasta aquí los principios. Como cualquier teoría, la TII se sostiene o se derrumba en función de que sus predicciones sean contrastables. La tesis principal de la teoría es que el nivel de conciencia de un sistema viene

dado por su Φ. Para testar eso, hace falta medir el Φ de los sistemas reales y ahí es donde comienza el problema. Resulta que medir el Φ es extraordinariamente dificultoso y, en la mayoría de casos, práctica o totalmente imposible. La principal razón es la manera inusual que tiene la TII de tratar la «información».

El uso convencional de la información en matemáticas, desarrollado por Claude Shannon en la década de 1950, es *relativo al observador*. La información relativa al observador (o *extrínseca)* consiste en el grado en que se reduce la incertidumbre, desde la perspectiva de un observador, al observar un sistema en un estado concreto. Por ejemplo, imaginemos que tiramos un solo dado muchas veces. Cada vez, observamos un resultado de seis posibles; cada vez, descartamos cinco opciones alternativas. Esto se corresponde con una determinada reducción de la incertidumbre (medida en bits) y es información «para» el observador.

Para medir la información relativa al observador, suele bastar con observar cómo se comporta un sistema a lo largo de cierto período de tiempo. Con los dados, podemos simplemente ir anotando el resultado que obtenemos con cada nueva tirada y esto nos permitirá calcular cuánta información se genera al salir cada número concreto. Si el sistema es una red de neuronas, bastará con registrar la actividad de las neuronas a lo largo del tiempo. Un observador externo puede ir registrando todos los estados diferentes en los que entran las neuronas, calcular las probabilidades asociadas a cada estado y luego medir la reducción de incertidumbre asociada al hecho de que la red esté en cualquiera de esos estados.[12]

Para la TII, sin embargo, la información no puede tratarse como algo relativo al observador. Esto se debe a que, para la TII, la información –información *integrada,* Φ– *es* en realidad conciencia, por lo que si tratáramos la información como algo relativo al observador, la propia conciencia sería también relativa al observador. Pero no lo es. El hecho de si yo soy consciente no debería depender (ni depende) de cómo tú o cualquier otra persona midáis mi cerebro.

En la TII, pues, la información debe tratarse como algo *intrínseco* a un sistema y no relativo a un observador externo. Debe definirse de tal modo que no dependa de nadie que observe desde fuera. Debe ser información «para» el propio sistema, y no para nadie ni nada más. Si no, la relación de identidad entre Φ y la conciencia, que conforma el núcleo central de la TII, no se sostiene.

Para medir información intrínseca, no basta solamente con observar cómo se comporta el sistema a lo largo del tiempo. Uno –como científico, como observador externo– tiene que saber todos los modos diferentes en que un sistema *podría* comportarse, incluso aunque nunca se comporte realmente de todos esos modos. Hay que diferenciar, pues, entre saber lo que un sistema *de verdad hace* a lo largo del tiempo (lo cual es fácil, al menos en principio, y es relativo al observador) y saber qué *podría hacer* un sistema aun cuando no lo haga nunca (lo cual suele ser difícil, si no imposible, pero es independiente del observador).

Si usamos el vocabulario de la teoría de la información, diremos que la diferencia entre esas dos situaciones es la misma que existe entre la distribución «empírica» de los estados de un sistema y su distribución de «máxima entropía» (así llamada porque refleja el nivel máximo de incertidumbre en torno a un sistema). Imagina que tiras dos dados varias veces. Puede que te salga un siete, un ocho y un once, además de otros números, pero que nunca te haya salido un doce. En esa situación, la distribución empírica no contendría ningún doce, pero la de máxima entropía sí, porque *podría haber salido* un doce en alguna tirada, aun cuando en esa secuencia concreta no haya sido así. Eso significa que cualquier resultado concreto –ya sea un siete, un ocho o un once– generaría más información con respecto a la distribución de máxima entropía (en la que se incluye el doce) que con respecto a la distribución empírica (que no lo incluye).

En comparación con la medición de la distribución empírica de un sistema mediante la simple observación del mismo a lo largo del tiempo, medir la distribución de máxima entropía es, en general, algo muy

difícil de llevar a cabo. Existen dos posibles maneras de proceder. La primera consiste en perturbar el sistema de todas las formas posibles y ver qué sucede, como un niño que pulsa todos los botones de un juguete nuevo para ver qué cosas es capaz de hacer. La segunda consiste en inferir la distribución de máxima entropía a partir de un conocimiento exhaustivo y completo del mecanismo físico del sistema, su «estructura de causas y efectos». Si lo sabes todo acerca de un mecanismo, a veces es posible conocer también todas las cosas que podría hacer, aun si en la práctica no las hace.[13] Si yo sé que un dado tiene seis caras, puedo deducir que dos dados pueden generar todos los números del dos al doce sin tener que hacer una sola tirada para averiguarlo.

Por desgracia, a menudo lo único a lo que podemos acceder es a la dinámica de un sistema, a lo que *hace* un sistema, más que a lo que un sistema *podría hacer*. Esto desde luego es así en el caso de los cerebros. Yo puedo registrar lo que hace tu cerebro con diferentes niveles de detalle, pero no tengo modo alguno de conocer su estructura física completa, ni puedo perturbar su actividad de todas las formas posibles. Por todos esos motivos, la tesis más característica y diferencial de la TII –la de que Φ *es* realmente la conciencia– es también la menos contrastable.

Hay más dificultades a las que se enfrentan los intentos de medir Φ, sea cual sea el tipo de información que elijamos. Una es que medir un Φ cualquiera obliga a hallar el modo apropiado de trocear el sistema para proceder a la mejor comparación posible entre el «todo» y las «partes».[14] En algunos sistemas –los cerebros divididos, por ejemplo–, esto es bastante sencillo (basta con seccionar por la mitad), pero, en general, se trata de un problema muy difícil, ya que el número de maneras posibles de subdividir un sistema crece exponencialmente a medida que lo hace su tamaño.

También hay cuestiones –más fundamentales incluso– referidas a qué se puede considerar un sistema y qué no, ya de entrada. ¿Cuál es el nivel de detalle correcto del espacio y del tiempo sobre el que podemos

calcular Φ? ¿Hablamos de neuronas y milisegundos, o de átomos y femtosegundos? ¿Podría estar consciente un país entero? Y, en ese caso, ¿podría un país estar más consciente que otro?[15] ¿Podríamos llegar incluso a considerar las interacciones entre las placas tectónicas a lo largo de escalas temporales geológicas como algo que integra información a escala planetaria?

Es importante aclarar que estos retos –incluido el de la medición de información intrínseca en vez de información extrínseca (o relativa al observador)– solo son problemas realmente para nosotros, los científicos, como observadores externos que tratamos de calcular Φ. Según la TII, todo sistema particular *tendría* simplemente un Φ. Haría lo que hace integrando información exactamente igual que cuando tú lanzas una piedra y esta dibuja un trayectoria en arco por el cielo sin que te haga falta calcularla con arreglo a las leyes de la gravedad. Que una teoría sea difícil de testar no significa que sea incorrecta: solo significa que es difícil de testar.

Dejemos a un lado el desafío que representa medir Φ y preguntémonos qué implicaciones tiene la TII si la teoría es correcta. Pues bien, resulta que seguir la lógica de la TII hasta el final nos conduce a unas consecuencias muy extrañas.

Imagínate que te abro el cráneo y acoplo a tu cerebro un puñado de neuronas nuevas, y que cada una de ellas se integra de una manera específica en el cableado de tu materia gris ya existente. Imagina, además, que, durante tu actividad diaria, esas nuevas neuronas nunca hacen nada en realidad. Pase lo que pase, hagas lo que hagas o veas lo que veas, no se activan en ningún momento. A todos los efectos, tu cerebro recién aumentado parece ser el mismo que el de antes. Pero –y he ahí la cuestión– tus nuevas neuronas están organizadas de tal forma que *podrían* activarse si el resto de tu cerebro se encontrase en cierto estado particular en el que, en la práctica, nunca se encuentra.

Digamos, por ejemplo, que estas neuronas nuevas solo se activarían si tú te llegaras a comer algún día una sandía *densuke,* una exótica fruta

que solo se encuentra en la isla japonesa de Hokkaido. Aun dando por supuesto que tú nunca te comes una *densuke* y que, por lo tanto, esas nuevas neuronas jamás se activan, la TII predice que *todas* tus experiencias conscientes cambiarían, aunque fuese de un modo muy sutil. Y lo harían porque ahora existen más estados potenciales en los que tu cerebro podría estar –es decir, en los que las nuevas neuronas podrían activarse–, y por lo tanto, Φ debe cambiar también.

El reverso de esa situación nos conduce a una predicción igualmente peculiar. Imagínate un grupo de neuronas que permanecen quietas en lo más hondo de tu córtex visual. Pese a estar interconectadas con otras –y, por consiguiente, pese a conservar su potencial de activarse si les llegan los *inputs* adecuados–, estas neuronas están ahí sin hacer nada. Un día, sin embargo, y gracias a cierta intervención experimental muy inteligente, alguien logra impedir que se vuelvan a activar de nuevo: pasan a estar *inactivadas,* más que *inactivas.* Aunque la actividad total del cerebro no habría cambiado lo más mínimo, la TII volvería a predecir un cambio en la experiencia consciente, pues el cerebro tendría entonces menos estados potenciales en los que entrar.

Aunque parezca asombroso, es posible que exista pronto una versión disponible de un experimento como ese gracias a la novedosa tecnología de la optogenética, que permite a los investigadores controlar con exquisito detalle la actividad de ciertas neuronas seleccionadas con precisión. La optogenética emplea técnicas genéticas para modificar neuronas específicas a fin de que se vuelvan sensibles a la luz de unas determinadas longitudes de onda. Luego, valiéndose de láseres o de módulos LED para proyectar luz sobre los cerebros de animales modificados genéticamente, los experimentadores pueden encender o apagar dichas neuronas.[16] En principio, la optogenética se podría utilizar para *inactivar* neuronas *ya inactivas,* y evaluar a partir de ahí los efectos sobre la percepción consciente, si los hubiera. No es un experimento sencillo y no nos proporciona un modo de medir Φ. Pero toda posibilidad de testar cualquier aspecto de la TII es emocionante, y

he tenido la suerte de participar en debates recientes –con Giulio Tononi y otros– orientados a conseguir que ese tipo de prueba pueda llevarse realmente a efecto.[17]

Vista desde un plano más general, otra de las rarezas de la TII es que su contundente premisa de que el Φ *es* la conciencia también implica que la *información en sí* existe, es decir, que posee cierto estatus ontológico definido en nuestro universo como el de la masa/energía y la carga eléctrica. (La ontología es el estudio de «lo que existe»). En cierto sentido, esto está en sintonía con el enfoque llamado en inglés *it from bit* (traducible como «a partir del bit»), del físico John Wheeler, probablemente el más famoso defensor de la idea de que todo lo que existe deriva en última instancia de la información: la información es lo primario, y todo lo demás se sigue de ella.[18]

Y esto nos conduce a una última implicación provocadora: el *pampsiquismo*. Siempre que se cuente con el mecanismo adecuado, es decir, con la forma adecuada de estructura de causas y efectos en un sistema, habrá un Φ diferente de cero y, por consiguiente, habrá conciencia. El pampsiquismo de la TII es una variante restringida de pampsiquismo, no como aquel otro que postula que la conciencia se halla extendida por la totalidad del universo cual fina capa de mermelada sobre una tostada. El de la TII nos dice, más bien, que la conciencia se encuentra allí donde haya información integrada (Φ). Esto puede ocurrir en lugares concretos (aquí o allí), pero no en todas partes a la vez.

La TII es original, ambiciosa y exuberante desde el punto de vista intelectual. Sigue siendo la única teoría neurocientífica que trata de dar respuesta seriamente al problema difícil de la conciencia. La TII es también extravagante, sin lugar a dudas, pero que algo sea raro no significa que esté equivocado. Casi toda la física moderna es extraña y, al mismo tiempo, menos errónea que la física del pasado. Pero el éxito de aquellas partes de la física moderna que hoy se han consolidado como menos equivocadas tiene mucho que ver con el hecho de que son

experimentalmente contrastables. Y he ahí el problema de la TII. Su audacia se cobra en ella un precio insoportablemente elevado: su tesis fundamental –la equivalencia entre Φ y nivel consciente– tal vez sea imposible de contrastar empíricamente.

Si tuviera que apostarme mi dinero, yo diría que la mejor manera de avanzar desde donde estamos es conservando el concepto central de la TII (el de que las experiencias conscientes son tanto informativas como integradas), pero renunciando al mismo tiempo a la idea de que Φ es a la conciencia lo que la energía cinética molecular media es a la temperatura. De ese modo, las conclusiones de la TII sobre la estructura de las experiencias conscientes se pueden realinear con la perspectiva del problema real. La adopción de este enfoque abre la puerta a desarrollar versiones alternativas de Φ, con aplicaciones prácticas: indicadores que terminen teniendo mucho en común con los indicadores de la complejidad que conocimos al final del capítulo anterior.

Mis colegas Adam Barrett y Pedro Mediano y yo llevamos años siguiendo esta estrategia. Hemos desarrollado varias versiones de Φ que funcionan con información relativa al observador, en vez de con información intrínseca. Esto nos permite medir Φ basándonos en el comportamiento observable de un sistema a lo largo del tiempo, sin preocuparnos de lo que podría hacer pero nunca hace. Tal como están las cosas actualmente, nuestras diversas versiones de Φ se comportan de maneras muy distintas, incluso con sistemas modelo muy simples.[19] Esto quiere decir que todavía queda mucho por hacer en el desarrollo de versiones de Φ que funcionen en la práctica, y que (esperemos) adquieran su aplicabilidad empírica gracias a (y no a pesar de) estar basadas en principios teóricos. Desde nuestra perspectiva, eso significa tratar la «integración» y la «información» como propiedades generales de las experiencias conscientes que hay que explicar, y no como tesis axiomáticas sobre lo que la conciencia «es» en realidad. Dicho de otro modo, se trata de abordar la conciencia asumiendo que esta se parece más a la vida que a la temperatura.

Nuestro periplo por los niveles de conciencia nos ha llevado por la inconsciencia de la anestesia y el coma, las apartadas regiones de los estados vegetativo y de mínima conciencia, los desconectados mundos del dormir y el soñar, las soleadas riberas de la plena consciencia despierta, y por lugares más alejados incluso, como los de la extraña hiperrealidad de la psicodelia. Lo que interconecta todos esos niveles es la idea de que cada experiencia consciente es tanto informativa como integrada, pues habita en el complejo terreno intermedio entre el orden y el desorden. Esta idea central de base ha dado origen a nuevos indicadores, como el ICP, que son útiles para la práctica y, al mismo tiempo, pueden servir para tender puentes explicativos –muy al estilo del enfoque del problema real– entre lo físico y lo fenoménico. Con la TII hemos entrado en uno de los más fascinantes y controvertidos territorios de la ciencia de la conciencia, donde la audacia topa con los límites de lo contrastable, y donde posiblemente se viene abajo la analogía entre la conciencia y la temperatura. Y aunque soy escéptico ante las tesis más generales de esta provocadora teoría, sigo igual de ansioso que hace años –cuando me comí aquel helado en compañía de Giulio Tononi– por ver cómo se desarrolla.[20]

Si lo pienso bien, Las Vegas era el sitio perfecto para debatir sobre la TII. ¿La información es real? ¿La conciencia está en todas partes? En Las Vegas cuesta creer que nada sea real más allá de la cruda sensación de la experiencia en sí. Incluso ahora, años después, puedo imaginarme allí, en el perpetuo anochecer del Venetian, con sus góndolas falsas trazando sus recorridos cual mecanismos de relojería. Soy sin duda consciente, pero *¿de qué?* En el Venetian, uno se siente tentado a pensar que todo es una especie de alucinación.

Como estamos a punto de ver, algo de verdad hay en tan peculiar idea.

II. CONTENIDO

4. PERCIBIR DE DENTRO HACIA FUERA

Abro los ojos y se me aparece un mundo. Estoy sentado en la terraza de la cubierta de una casa de madera muy desvencijada, en lo alto de un bosque de cipreses, unos kilómetros al norte de Santa Cruz (California). Es por la mañana temprano. Si miro fuera, puedo ver árboles altos envueltos todavía en la fresca niebla oceánica que cae por allí todas las noches y hace que la temperatura se desplome. No puedo ver el suelo: es como si la terraza, los árboles y yo flotáramos todos juntos en la bruma. Hay unas viejas sillas de plástico –estoy sentado en una de ellas–, una mesa y una bandeja donde han dispuesto café y pan. Puedo oír el canto de los pájaros, susurros por detrás –de las personas con las que estoy– y un murmullo distante de algo que no logro identificar. No todas las mañanas son así; esta es una buena mañana. Trato de convencerme (y no es la primera vez) de que este extraordinario mundo es una construcción de mi cerebro, una especie de «alucinación controlada».

Siempre que estamos conscientes, somos conscientes de algo, o de muchas cosas. Son los *contenidos* de la conciencia. Para entender cómo se producen y qué quiero decir cuando hablo de una alucinación controlada, vamos a cambiar de perspectiva. Imagina, por un momento, que eres un cerebro.

Esfuérzate un poco e intenta pensar cómo es estar ahí, sellado dentro de la huesuda cámara acorazada del cráneo, tratando de comprender lo que hay ahí fuera, en el mundo. No hay luz, ni sonido, ni nada: estás en la más completa oscuridad y el más absoluto silencio. Cuando intenta formarse percepciones, lo único con lo que cuenta el cerebro es una

andanada constante de señales eléctricas que solo guardan una relación indirecta con las cosas de fuera, del mundo, sean las que sean. Estos *inputs* sensoriales no vienen con etiquetas incorporadas («soy de una taza de café», «soy de un árbol»). Ni siquiera llegan con etiquetas que anuncien su modalidad, es decir, si son visuales, auditivos, táctiles o de otros tipos con los que estamos menos familiarizados, como la termocepción (la sensación de temperatura) o propiocepción (la sensación de la posición corporal).*

¿Cómo transforma el cerebro estas señales sensoriales inherentemente ambiguas en un mundo perceptual coherente, lleno de objetos, personas y lugares? En la segunda parte de este libro, exploraremos la idea según la cual el cerebro es una «máquina de predicción» y lo que vemos, oímos y sentimos no es más que su «mejor conjetura» acerca de las causas de los *inputs* sensoriales que recibe. Siguiendo la lógica de esa idea hasta el final, veremos que los contenidos de la conciencia son una especie de sueño despierto –una alucinación controlada– que es más y (también) menos que lo que el mundo real es de verdad.

A continuación, hablaré de una perspectiva de la percepción que muchos considerarán de sentido común. Llamémosla la perspectiva del «cómo parece que son las cosas».

Ahí fuera hay una realidad independiente de la mente, llena de objetos, personas y lugares que poseen realmente propiedades como color, forma, textura, etcétera. Nuestros sentidos funcionan como ventanas transparentes a ese mundo, detectan esos objetos y sus rasgos, y transmiten esa información al cerebro, donde unos complejos procesos neuronales la leen y forman con ella percepciones. Una taza de café que está ahí fuera, en el mundo, conduce a que tengamos la percepción de una taza de café generada dentro del cerebro. En cuanto a quién o qué es el agente perceptor, ¿acaso no es el «yo»: ese yo que está «detrás de los ojos», podría decirse, que es el receptor de oleada tras oleada de datos sensoriales, y que usa sus lecturas perceptivas para guiar la conducta, para decidir qué hacer a continuación? Ahí hay una taza de

café. La percibo y la cojo. Tengo una sensación, pienso y luego actúo.

El concepto en sí nos resulta atractivo. Ciertos patrones de pensamiento consolidados a lo largo de décadas, o tal vez de siglos, nos han acostumbrado a la idea de que el cerebro es una especie de ordenador instalado en el interior del cráneo que procesa información sensorial para construir una imagen interna del mundo externo en beneficio del yo. Tan familiar nos resulta esta imagen que puede dificultarnos el concebir cualquier alternativa razonable. De hecho, muchos neurocientíficos y psicólogos continúan entendiendo así la percepción, es decir, como un proceso de detección «ascendente» (de abajo arriba, o *bottom up)* de características.

Así se supone que funciona la percepción concebida como un proceso ascendente. El mundo proyecta unos estímulos –ondas de luz y de sonido, moléculas que transmiten sabores, olores, etcétera– que impactan en los órganos sensoriales y provocan unos impulsos eléctricos que fluyen «hacia arriba» (o «hacia dentro») en el cerebro. Estas señales sensoriales pasan por varias etapas de procesamiento diferenciadas, representadas por las flechas negras de la imagen de más abajo; en cada una de esas etapas, se analizan características cada vez más complejas. Pongamos el ejemplo de la visión. En las primeras fases, la vista tal vez responda a factores como la luminancia o los bordes, y en otras etapas posteriores, y más profundas, a partes de objetos, como ojos y orejas, o ruedas y espejos retrovisores. Y en etapas más avanzadas todavía, puede que reaccione a objetos enteros, o a categorías de objetos, como rostros, coches y tazas de café.

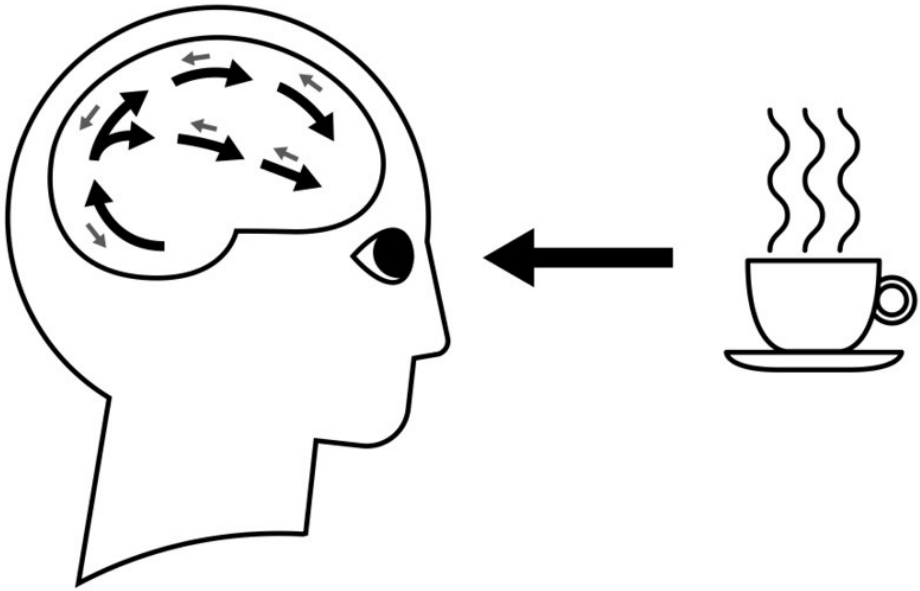

Fig. 3. La percepción concebida como una detección ascendente (*bottom-up*) de características

De este modo, el mundo externo, con sus objetos, sus personas y sus categorías de todo tipo, queda condensado en una serie de rasgos extraídos del torrente de datos sensoriales que afluye al cerebro de una forma parecida a las capturas de unos pescadores que fueran pescando piezas de tamaño y complejidad crecientes cuanto más avanzan por un río. Las señales que fluyen en el sentido contrario –«de arriba abajo» o «de dentro hacia fuera», que serían las pequeñas flechas grises de la imagen– solo sirven para depurar o, en el peor de los casos, obstaculizar el importantísimo flujo de información sensorial que va de abajo arriba.

A simple vista, esta concepción de la percepción como proceso ascendente parece encajar bien con lo que sabemos de la anatomía del cerebro. En la corteza cerebral, cada modalidad perceptiva se asocia a unas regiones específicas: hay un córtex visual, un córtex auditivo, etcétera. Dentro de cada región, el procesamiento perceptivo está

organizado de manera jerárquica. En el sistema visual, los niveles inferiores –como, por ejemplo, el córtex visual primario– están más próximos a los *inputs* sensoriales, mientras que los niveles superiores –como el córtex inferotemporal– están varias fases de procesamiento más allá. En términos de conectividad, las señales de cada nivel se agrupan en el nivel inmediatamente superior, de manera que las neuronas de esos niveles más avanzados pueden reaccionar a características que tal vez estén repartidas por el espacio o el tiempo, como cabría esperar.[1]

También los estudios de la actividad cerebral parecen cuadrar bien con la perspectiva ascendente. En experimentos cuyos orígenes se remontan a varias décadas atrás –cuando comenzaron investigándose los sistemas visuales de los gatos y los monos–, se ha mostrado reiteradamente que las neuronas de las etapas tempranas (inferiores) del procesamiento visual responden a rasgos simples como los bordes, mientras que las neuronas de fases posteriores (superiores) reaccionan a características complejas como las caras. En experimentos más recientes, con técnicas de neuroimagen como la IRMf, se ha revelado algo muy parecido en cerebros humanos.[2]

Incluso se pueden construir «sistemas perceptores» artificiales de ese modo, al menos, de carácter rudimentario. La clásica teoría computacional de la visión que el científico informático David Marr postuló en 1982 es, a un tiempo, una referencia estándar para la concepción ascendente de la percepción y un libro de recetas muy práctico para el diseño y fabricación de sistemas de visión artificial.[3] Algunos sistemas de visión mecánica más recientes que incorporan redes neuronales artificiales –como las redes de «aprendizaje profundo»– están alcanzando actualmente unos impresionantes niveles de rendimiento que, en algunas situaciones, son incluso comparables a las capacidades de los seres humanos.[4] También estos sistemas se basan a menudo en teorías del enfoque ascendente.

Con todos estos puntos a su favor, se diría que la perspectiva de la percepción entendida como un proceso ascendente, muy fundada en la

«apariencia» del funcionamiento de las cosas, continúa pisando terreno sólido hoy en día.

> LUDWIG WITTGENSTEIN: ¿Por qué dice la gente que era natural pensar que el Sol giraba alrededor de la Tierra y no que la Tierra giraba sobre su eje?
>
> ELIZABETH ANSCOMBE: Supongo que porque parecía que el Sol giraba alrededor de la Tierra.
>
> LUDWIG WITTGENSTEIN: De acuerdo, pero ¿qué habría parecido si hubiese *parecido* que la Tierra giraba sobre su eje?

En este delicioso diálogo entre Wittgenstein y su colega filósofa (y biógrafa) Elizabeth Anscombe, el legendario pensador austríaco usa el ejemplo de la revolución copernicana para ilustrar la idea de que lo que las cosas *parecen* no es necesariamente lo que *son*.[5] Aunque *parezca* que es el Sol el que gira alrededor de la Tierra, es obvio que es el hecho de que la tierra gire sobre su propio eje lo que nos da la noche y el día, y es el Sol, y no la Tierra, el que está situado en el centro del Sistema Solar. Nada nuevo hay ahí, pensarás, y tienes razón. Pero Wittgenstein apuntaba a algo más profundo. Su verdadero mensaje para Anscombe era que, incluso con un mayor conocimiento sobre cómo son las cosas en realidad, sigue existiendo un nivel en el que las cosas no dejan de parecer lo que siempre han parecido. El sol sale por el este y se pone por el oeste, como siempre ha hecho.

Con la percepción, ocurre algo muy similar a esto que acabo de explicar sobre el sistema solar. Abro los ojos y *parece* que hay un mundo real ahí fuera. Hoy me encuentro en mi casa, en Brighton. No hay cipreses como los que había en Santa Cruz: solo el habitual batiburrillo de objetos sobre mi mesa de trabajo, una silla roja en el rincón y, fuera, del otro lado de la ventana, un sube y baja de chimeneas. Son objetos que *parecen poseer* formas y colores específicos, y los más cercanos, incluso olores y texturas. Así *parecen* las cosas.

Aunque pueda *parecer* que mis sentidos son como unas ventanas

transparentes que se me abren a una realidad independiente de la mente, y que la percepción es un proceso de interpretación o «lectura» de los datos sensoriales, lo que ocurre realmente es –según yo lo entiendo– muy distinto. Las percepciones no proceden de abajo arriba o de fuera hacia dentro, sino que van principalmente de arriba abajo, o de dentro hacia fuera. Lo que experimentamos está construido a partir de las predicciones (o «mejores conjeturas») que el cerebro hace acerca de las causas de las señales sensoriales. Como ocurriera con la revolución copernicana, esta concepción descendente de la percepción sigue siendo congruente con buena parte de la evidencia existente y deja inalterados muchos aspectos de lo que las cosas parecen, pero, al mismo tiempo, lo cambia todo.

Esta no es una idea completamente novedosa, ni mucho menos. Los primeros destellos de una teoría de la percepción concebida como un proceso descendente aparecen en la Antigua Grecia, con el famoso «mito de la caverna» de Platón. Unos presos, encadenados y encarados contra un muro vacío toda su vida, solo han alcanzado a ver el juego de sombras proyectado por los objetos que pasaban ante una hoguera que alumbra la estancia desde detrás de donde ellos están. Los presos ponen nombres a las sombras, porque, para ellos, las sombras son lo real. La alegoría del mito platónico viene a decirnos que nuestras propias percepciones conscientes son como esas sombras, reflejos indirectos de unas causas escondidas que jamás llegamos a ver directamente.

Más de mil años después, pero, aun así, hace mil años para nosotros, el erudito árabe Alhacén escribió que la percepción, en el aquí y el ahora, depende de unos procesos de «juicio e inferencia», más que de un acceso directo a una realidad objetiva.[6] Cientos de años más tarde, Immanuel Kant se dio cuenta de que el caos de los datos sensoriales ilimitados no tendría nunca sentido para nosotros si no lo estructuráramos con unas concepciones preexistentes, entre las que, según él, se incluían marcos apriorísticos como el espacio y el tiempo. El término kantiano noúmeno alude a «la cosa en sí» *(Ding an sich),* una

realidad independiente de la mente y oculta tras un velo sensorial que la hace eternamente inaccesible a la percepción humana.

La historia continúa, ya en el ámbito de la neurociencia, con el físico y fisiólogo alemán Hermann von Helmholtz. A finales del siglo XIX, entre sus muchas contribuciones de peso, Helmholtz propuso la idea de que la percepción era un proceso de «inferencia inconsciente». Los contenidos de la percepción, argumentó, no vienen dados por las señales sensoriales en sí, sino que tienen que ser *inferidos* mediante la combinación de esas señales con las expectativas o creencias que el cerebro se forma acerca de sus causas. Al calificar ese proceso de inconsciente, Helmholtz demostró haber entendido que no somos conscientes de los mecanismos por los que se producen las inferencias perceptivas, sino solo de sus resultados. Los juicios perceptivos –sus «inferencias inconscientes»– van rastreando sus causas en el mundo mediante la actualización continua y activa de las mejores conjeturas perceptuales a medida que van llegando nuevos datos sensoriales. El propio Helmholtz consideró que lo que él estaba facilitando era una versión científica de la idea kantiana de que la percepción no nos permite conocer las cosas del mundo de forma directa: solo podemos inferir que las cosas están ahí, detrás del velo sensorial.[7]

La idea de la «percepción como inferencia», central para Helmholtz, ha tenido una considerable influencia y terminó adoptando muchas formas distintas a lo largo del siglo XX. En la década de los cincuenta, el movimiento de la «nueva mirada» *(new look)* en psicología puso el énfasis en cómo los factores sociales y culturales podían influir en la percepción. Por ejemplo, en un estudio muy difundido se descubrió que los niños de familias pobres sobreestimaban el tamaño de las monedas, mientras que los de familias acomodadas no lo hacían.[8] Por desgracia, muchos experimentos de ese tipo, pese a ser fascinantes, pecaban de una pobre ejecución conforme a los criterios metodológicos actuales, por lo que sus resultados no siempre eran de fiar.

En los años setenta, el psicólogo Richard Gregory se basó en las ideas

de Helmholtz para darles un enfoque distinto: usó su teoría de la percepción para convertirla en una especie de teoría neural fundada en la dinámica del «contraste de hipótesis». Según Gregory, del mismo modo que los científicos contrastan y actualizan las hipótesis científicas obteniendo datos de experimentos, el cerebro se dedica continuamente a formular hipótesis perceptivas acerca de cómo es el mundo (basándose en experiencias pasadas y en otras formas de información almacenada) y a contrastarlas adquiriendo datos de los órganos sensoriales. El contenido perceptivo, según Gregory, viene determinado por las hipótesis que el cerebro encuentra más fundamentadas.[9]

Tras haber adquirido cierta prominencia y haberla perdido en el medio siglo transcurrido desde entonces, la idea de la percepción como inferencia ha vuelto a cobrar un nuevo impulso en los últimos diez años, más o menos. Varias teorías nuevas han florecido bajo las categorías generales de «codificación predictiva» y «procesamiento predictivo». Aunque son teorías que difieren en cuanto a los detalles, comparten la premisa de que la percepción depende de algún tipo de inferencia de base cerebral.[10]

Lo que mejor capta mi enfoque de esa antigua idea de Helmholtz –y de sus manifestaciones contemporáneas– es la noción de la percepción entendida como una *alucinación controlada,* una expresión que oí por primera vez de boca del psicólogo británico Chris Frith muchos años atrás.[*] Los ingredientes esenciales de la alucinación controlada, tal como yo la concibo, son los siguientes.

En primer lugar, el cerebro está haciendo predicciones constantemente acerca de las causas de sus señales sensoriales, predicciones que caen en cascada, desde arriba y hacia abajo, a través de las jerarquías perceptivas del cerebro (las flechas grises de la imagen de la página siguiente). Si, por ejemplo, estás mirando una taza de café, tu córtex visual estará formulando predicciones acerca de las causas de las señales sensoriales que tienen su origen en dicho objeto.

En segundo lugar, las señales sensoriales –que entran en torrente en el

cerebro desde abajo hacia arriba, o de fuera hacia dentro– mantienen esas predicciones perceptivas útilmente conectadas a su causa (en este caso, una taza de café). Estas señales sirven de *errores de predicción* que registran la diferencia entre lo esperado por el cerebro y lo obtenido por él en cada nivel de procesamiento. Las predicciones generadas «de arriba abajo» se van ajustando entonces para suprimir los errores de predicción (que van «de abajo arriba») y, de ese modo, las «mejores conjeturas» del cerebro no pierden el contacto con sus causas en el mundo.[11] Desde ese punto de vista, la percepción se produce a través de un proceso continuo de *minimización del error de predicción.*

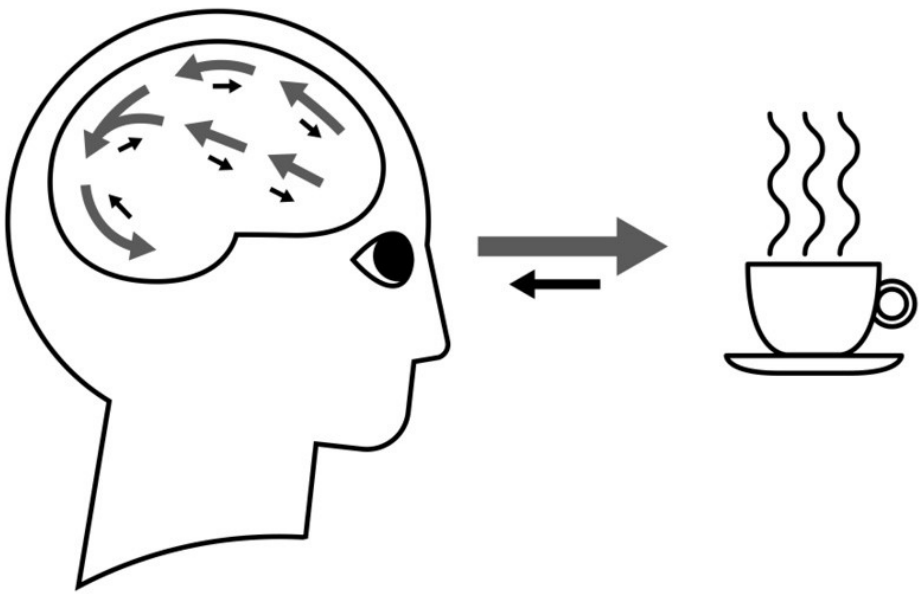

Fig. 4. La percepción concebida como una inferencia descendente (*top-down*)

El tercer (y más importante) ingrediente en el enfoque de la alucinación controlada es la tesis de que la experiencia perceptiva –en este caso, la experiencia subjetiva de «ver una taza de café»– está determinada por el contenido de las predicciones (que proceden de

arriba abajo), y no por las señales sensoriales (que proceden de abajo arriba). Jamás experimentamos señales sensoriales propiamente dichas, sino solo interpretaciones de las mismas.

Estos ingredientes, combinados, dan origen a un giro copernicano en nuestra manera de concebir la percepción. Puede *parecer* que el mundo se nos revela directamente a nuestras mentes conscientes a través de nuestros órganos sensoriales. Y, de hecho, con esa mentalidad, es natural que pensemos que la percepción es un proceso ascendente de detección de características: una «lectura» del mundo que nos rodea. Pero lo que percibimos en realidad es una fantasía neuronal producida de arriba abajo (o de dentro hacia fuera) refrenada por la realidad, no la imagen de lo que quiera que sea esa realidad llegada a nosotros a través de una ventana transparente.

Y volviendo a Wittgenstein, ¿qué parecería la realidad si *pareciera como si* la percepción fuese una mejor conjetura generada desde arriba y hacia abajo? Pues igual que el sol sigue saliendo por el este y poniéndose por el oeste, si todo *pareciera como si* la percepción fuese una alucinación controlada, la taza de café sobre la mesa (y la totalidad de la experiencia perceptiva de la persona) seguiría pareciendo igual que como siempre ha sido y siempre será.

Cuando pensamos en alucinaciones, solemos imaginarnos una especie de percepción generada a nivel interno: una visión o una audición de algo que no está realmente ahí fuera, como puede ocurrir en casos de esquizofrenia o durante aventuras psicodélicas como las de Albert Hofmann. Al hacer esas asociaciones, contraponemos la alucinación a la percepción «normal», que es la que se supone que refleja las cosas tal como estas existen en el mundo externo. Sin embargo, desde el enfoque de la percepción concebida como un proceso descendente, esa distinción tan drástica pasa a ser una cuestión de grado. Tanto la percepción «normal» como la alucinación «anómala» implican unas predicciones generadas a nivel interno a propósito de las causas de los *inputs* sensoriales, y ambas comparten un conjunto fundamental de

mecanismos cerebrales. La diferencia estriba en que, en la percepción «normal», lo que percibimos está ligado a *(controlado por)* las causas en el mundo, mientras que, en el caso de la alucinación, nuestras percepciones han perdido (hasta cierto punto) su contacto con esas causas. Cuando alucinamos, nuestras predicciones perceptivas no se actualizan adecuadamente en función de los errores de predicción.

En definitiva, si la percepción es alucinación controlada, entonces también la alucinación puede entenderse como percepción descontrolada. Son diferentes, pero preguntarnos dónde trazar la línea de separación es como preguntarnos dónde está la frontera entre el día y la noche.

Vamos a probar un poco la teoría de la alucinación controlada preguntándonos lo que significa experimentar perceptivamente el *color*.

Nuestro sistema visual, por asombroso que sea, solo responde a una estrecha franja del espectro electromagnético: la que va desde los mínimos del infrarrojo hasta los máximos del ultravioleta. Cada color que percibimos (y, de hecho, cada parte de la totalidad de cada uno de nuestros mundos visuales) se basa en esa fina rodaja de la realidad. Basta con saber esto para entender que la experiencia perceptiva no puede ser una representación exhaustiva de un mundo objetivo externo: es menos que eso…, pero es más que eso también.

Si se lo preguntas a un neurofisiólogo, probablemente te dirá que percibes un color particular cuando las células cono de tu retina se activan en una determinada proporción. Y no te estará mintiendo, pero esa no es, ni de lejos, la historia completa. No existe una correspondencia absoluta de uno a uno entre las actividades de las células sensibles al color (los conos) y la experiencia del color. El color que experimentas depende de una compleja interacción entre la luz reflejada desde una superficie y la iluminación general del entorno en el que tú te encuentres en ese momento. Para ser más precisos, depende de cómo tu cerebro elabore inferencias –sus «mejores conjeturas» en cada momento– acerca de cómo se produce esa interacción.

Pon al sol una hoja de papel blanco y seguirá viéndose blanca, pese a que la luz que reflejará entonces poseerá una composición espectral muy distinta, dadas las diferencias entre la luz solar (azulada) y la luz interior (amarillenta). Tu sistema visual compensa automáticamente esas diferencias en la iluminación ambiental –«descuenta el iluminante», como les gusta decir a los investigadores de la visión– para que tu experiencia del color capte una propiedad invariable de ese papel: el *modo en el que* el papel refleja la luz. El cerebro infiere esa propiedad invariante porque hace su mejor conjetura acerca de las causas de los *inputs* sensoriales –continuamente cambiantes– que recibe. La blancura es el aspecto fenomenológico de esa inferencia: es cómo las inferencias del cerebro sobre esa propiedad invariante aparecen en nuestra experiencia consciente.

Esto significa que el color no es una propiedad definida de las cosas en sí, sino que es, más bien, un mecanismo útil que ha encontrado la evolución para que el cerebro pueda reconocer los objetos y hacerles seguimiento bajo unas condiciones de iluminación cambiantes. Cuando tengo la experiencia subjetiva de ver la silla roja en el rincón de la habitación, eso no quiere decir que la silla *sea realmente roja,* porque ¿acaso significaría algo siquiera que una silla posea una propiedad fenomenológica como la rojez? Las sillas no son rojas del mismo modo que no son feas, anticuadas ni vanguardistas. En realidad, lo que sucede es que la superficie de la silla posee una propiedad particular, su *modo-de-reflejar-la-luz,* que mi cerebro registra a través de sus mecanismos de percepción. La rojez es el aspecto subjetivo, fenomenológico, de ese proceso.

¿Significa eso que la rojez de la silla ha pasado de estar «ahí fuera» en el mundo a estar «aquí dentro» del cerebro? En cierto sentido, la respuesta es un rotundo «no». No existe rojo alguno en el cerebro en el sentido ingenuo de que haya un pigmento rojizo –real o figurado– dentro de la cabeza que se pueda inspeccionar con una cámara de vídeo en miniatura que sirva sus imágenes a otro sistema visual que, a su vez,

tenga su propia minicámara dentro, y así sucesivamente. Suponer que una propiedad percibida del mundo exterior (la rojez) tiene que resustanciarse de algún modo en el cerebro para que se pueda producir la percepción es caer en la trampa de lo que el filósofo Daniel Dennett ha llamado la falacia de la «doble transducción». En concreto, la falacia está en suponer que la retina es una especie de transductor que transforma una «rojez» externa en unos patrones de actividad eléctrica que luego hay que reconstituir –transformar de nuevo mediante otro proceso de transducción– en una «rojez» interna. Como bien explica Dennett, esta forma de razonamiento no explica nada.[12] El único sentido en el que se podría localizar la rojez «en el cerebro» es simplemente en el de que es en ese órgano donde se encuentran los mecanismos que subyacen a la experiencia perceptiva. Y estos mecanismos, huelga decirlo, no son rojos.

Cuando miro una silla roja, la rojez que experimento con la vista depende tanto de unas propiedades de la silla como de unas propiedades de mi cerebro. Se corresponde con el contenido de un conjunto de predicciones perceptivas acerca de las formas en que un tipo específico de superficie refleja la luz. No hay una rojez-en-sí en el mundo ni en el cerebro. Como bien dijo Paul Cézanne, «el color es el lugar donde coinciden nuestro cerebro y el universo».[13]

La conclusión general aquí es que todo esto es también aplicable mucho más allá del ámbito de la experiencia cromática. Es válido para toda clase de percepción. El panorama multisensorial inmersivo de tu escena perceptiva particular, justo en este lugar y en este momento, es eso que te proporciona tu cerebro tratando de salir al encuentro del mundo, y tiene tanto de escritura como de lectura. La totalidad de la experiencia perceptiva es una fantasía neuronal que permanece uncida al mundo a través de la elaboración y la reelaboración continuas de mejores conjeturas perceptuales, de alucinaciones controladas.

Incluso podría decirse que todos estamos alucinando todo el tiempo. Sucede simplemente que cuando nos ponemos de acuerdo sobre nuestras

alucinaciones, encontramos algo a lo que llamar realidad.[14]
Consideremos tres ejemplos de cómo las expectativas perceptivas moldean la experiencia consciente, ejemplos que puedes experimentar por ti mismo.

Si estabas conectado a las redes sociales, aunque fuera indirectamente, o leíste algún periódico durante una semana en concreto de febrero de 2015, seguro que recuerdas *El vestido*. La mañana del miércoles de aquella semana, al llegar a mi despacho, me encontré con un alud de correos electrónicos y mensajes de voz. Acabábamos de publicar un librito sobre las ilusiones visuales, del que era coautor,[15] y los medios andaban como locos buscándole explicaciones a un fenómeno que, de repente, se había propagado por todo internet. *El vestido* (#TheDress) es una foto fortuita en la que se muestra un vestido que a algunos internautas les pareció azul y negro, y a otros, blanco y dorado.[*] Quienes lo veían de un modo estaban tan convencidos de que tenían razón que no se podían creer que nadie lo viera de otra forma; así que el volcán de internet entró en erupción con afirmaciones, réplicas y contrarréplicas.

Al principio, pensé que tal vez se trataba de un engaño. A mí me pareció tan evidente que el vestido era azul y negro –a mí y a las cuatro primeras personas del laboratorio a las que se lo mostré– que sentí una mezcla de alivio y de sorpresa cuando la quinta dijo que le parecía blanco y dorado. Al final, resultó que –como en el mundo en general– aproximadamente la mitad del personal del laboratorio se decantó por la opción azul-negro y la otra mitad, por la blanco-dorado.

Una hora más tarde, aparecí en la BBC tratando de explicar lo que estaba ocurriendo. Una opinión de consenso parecía imponerse en aquel momento: el efecto tenía que ver con nuestra tendencia como observadores a descontar la fuente iluminante, es decir, con el proceso de tomar en consideración la luz ambiental a la hora de percibir los colores. La idea era que ese proceso podía funcionar de forma distinta para unas personas y otras, y que podía hacerlo hasta un punto que

normalmente no se manifestaba con tanta claridad y del que no se había tenido constancia hasta entonces, pero que justamente era el que marcaba las diferencias en aquel asunto del vestido.

Enseguida hubo quienes señalaron que la fotografía del vestido estaba sobreexpuesta y descontextualizada (es la propia prenda la que ocupa la mayor parte de la imagen), y eso podía jugar con el modo en que el cerebro genera la percepción del color a partir del contexto. Si, por alguna razón, tu sistema visual está acostumbrado a una luz ambiental amarillenta –tal vez porque pasas mucho tiempo en espacios interiores–, entonces es posible que sea más propenso a inferir una combinación azul-negro porque presupone una fuente iluminante tendente al amarillo. Si, por el contrario, eres la típica persona de vida saludable, acostumbrada a los espacios abiertos y, por lo tanto, con un córtex visual frecuentemente empapado de luz solar azulada, tal vez tiendas a ver una combinación blanco-dorado.

Enseguida comenzaron a hacerse toda clase de experimentos con aquella foto: observarla fijamente en una habitación poco iluminada y, acto seguido, salir rápidamente afuera, a observarla a plena luz del día; correlacionar la prevalencia de informes (testimonios) individuales de percepción de blanco-dorado con los promedios de horas de sol según los países; estudiar si las personas mayores tienden más a ver la combinación azul-negro que las jóvenes. Pronto, todo un sector de actividad parecía haber nacido en torno a la comprobación de estas y un sinfín más de hipótesis.[16]

El hecho de que las personas tengan experiencias tan diferentes –e informen de ellas con tanta seguridad– a propósito de exactamente la misma imagen es una prueba bastante convincente de que nuestras experiencias perceptivas del mundo son construcciones internas moldeadas por las idiosincrasias de nuestra biología y nuestra historia personales. La mayor parte del tiempo, presuponemos que cada individuo ve el mundo más o menos de la misma manera, y puede que así sea la mayor parte del tiempo. Pero aun si ese es el caso, no lo es

porque las sillas rojas sean *rojas de verdad:* se necesita una situación tan inusual como la del vestido para detectar mejor las pequeñas diferencias en cuanto a las maneras que nuestros respectivos cerebros tienen de decidir sus propias mejores conjeturas perceptivas posibles.

El segundo ejemplo es una ilusión visual muy apreciada llamada Tablero de Ajedrez de Adelson. Es un ejemplo que nos muestra que la influencia de las predicciones en la percepción no se limita a situaciones extrañas como la del vestido, sino que ocurre en todos los momentos y lugares. En la imagen de la página siguiente, fíjate en el tablero de ajedrez de la izquierda y compara los escaques A y B. A lo mejor, notas que A te parece más oscuro que B. A mí me lo parece y lo mismo le sucede a todo el mundo al que se lo he enseñado a lo largo de los años. Nunca he percibido ninguna diferencia entre individuos en este caso.

En realidad, A y B tienen la misma tonalidad exacta de gris. El tablero de la derecha nos lo demuestra al unir ambas casillas mediante un rectángulo que tiene ese mismo tono exacto de gris, aun cuando en el tablero de izquierda continúan pareciéndonos distintas. Saber que son iguales no nos ayuda a cambiar esa impresión. He mirado detenidamente estas imágenes miles de veces, y persiste en mí, tozuda, la impresión de que A y B (las de la izquierda) están sombreadas con dos tonos de gris diferentes.*

Lo que pasa en este caso es que la percepción de la grisura viene determinada no por las ondas luminosas reales que nos llegan de A o de B –que son idénticas–, sino por la mejor conjetura posible que se hace el cerebro acerca de qué ha causado esas combinaciones concretas de ondas luminosas, y –como en el caso del vestido– eso depende del contexto. B está a la sombra y A no lo está, y el sistema visual del cerebro tiene incrustado muy a fondo en sus circuitos el conocimiento de que los objetos ensombrecidos parecen más oscuros. Exactamente del mismo modo que el cerebro ajusta sus inferencias perceptivas basándose en la iluminación ambiental, también ajusta las que hace a propósito del tono de B sobre la base de su conocimiento anterior acerca de las

sombras. De ahí que, en el tablero de ajedrez de la izquierda, percibamos B más claro que A (situado fuera de la sombra). Sin embargo, en el tablero de la derecha, el contexto que nos proporciona la sombra se ve interrumpido por las dos barras grises superpuestas, lo que nos permite ver que A y B son, en realidad, idénticos.

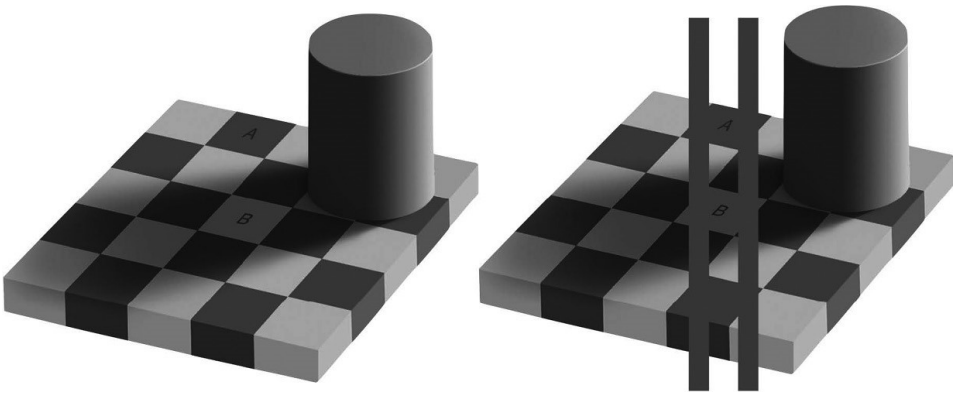

Fig. 5. Tablero de Ajedrez de Adelson

Todo esto es completamente automático. No eres consciente (o, por lo menos, no lo eras) de que tu cerebro posee y usa unas expectativas previas referidas a las sombras cuando elabora sus predicciones perceptivas. Tampoco es que esté fallando el sistema visual. Un sistema visual útil no es aquel que funciona como un fotómetro al estilo de los que usan los fotógrafos. La función de la percepción, al menos en un sentido aproximado, consiste en averiguar las causas más probables de las señales sensoriales, no en proporcionarnos una constancia de las señales sensoriales en sí, sea lo que sea que eso signifique.

Por último, este otro ejemplo nos revela lo rápido que las nuevas predicciones pueden influir en la percepción consciente. Fíjate en la imagen de más abajo. Lo más probable es que lo único que aciertes a ver en ella sea un amasijo de manchas negras y blancas. Luego, cuando ya hayas leído el resto de esta frase, echa un vistazo a la imagen que aparece en la página 120, y vuelve de nuevo aquí.

Fig. 6. ¿Qué es esto?

De acuerdo, ya estás de vuelta. Pues ahora vuelve a mirar la imagen de esta página: debería de parecerte bastante diferente. Allí donde antes solo había un desorden de manchas, ahora hay objetos distinguibles, hay *cosas* y algo que está ocurriendo. Esta es una imagen «bitonal» o de «test de Mooney». En cuanto la has percibido, es difícil de «despercibir». Las imágenes bitonales se crean tomando una foto, pasándola a escala de grises y aplicándole un minucioso efecto umbral para reducirla a extremos de negro y blanco –los «dos tonos»– en los que se pierden los detalles intermedios. Si se hace bien (y con la fotografía adecuada), se vuelve muy difícil averiguar lo que se representa en la imagen. Eso… hasta que ves el original, en cuyo caso, la imagen bitonal de pronto se transforma en una escena coherente.

Lo asombroso de este ejemplo es que, cuando vuelves a mirar la imagen bitonal original ahora, las señales sensoriales que llegan a tus ojos no han cambiado en absoluto respecto a las de la primera vez que

la viste. Lo único que ha variado son las predicciones que elabora tu cerebro acerca de las causas de esos datos sensoriales, y eso transforma lo que estás viendo de forma consciente.[17]

Este fenómeno no es exclusivo de la visión. Existen también ejemplos auditivos muy convincentes, como el «habla de onda sinusoidal».[18] En este caso, se procesa una frase hablada suprimiendo todas las altas frecuencias que hacen que el discurso oral normal sea comprensible. El resultado suele sonar como un silbido ruidoso carente de todo sentido: el equivalente auditivo de una imagen bitonal. Luego, la persona escucha la frase original sin procesar, y cuando vuelve a oír la versión «sinusoidal» de nuevo, de pronto todo se le vuelve nítido, porque aquí, igual que ocurre con las imágenes bitonales, el hecho de tener una predicción fuerte acerca de las causas de las señales sensoriales cambia –enriquece– la experiencia perceptiva.

Tomados en conjunto, estos ejemplos (aun reconociendo que son deliberadamente simples) revelan el carácter generativo y creativo de la percepción como acto en el que se interpretan proactiva y contextualmente unas señales sensoriales con las que al mismo tiempo se produce una interacción. Y como ya he mencionado antes, el principio según el cual la experiencia perceptiva se construye a partir de predicciones de base cerebral es aplicable de forma generalizada, no ya a la vista y el oído, sino a todas nuestras percepciones y en todo momento.[19]

Una implicación importante de ese principio es que nunca experimentamos el mundo «como es». De hecho, como bien señaló Kant con su concepto del noúmeno, es difícil saber qué significaría algo semejante. Incluso algo tan básico como el color, como hemos visto, solo existe en la interacción entre un mundo y una mente. Así que, aunque tal vez nos sorprenda descubrir que las ilusiones perceptivas –del tipo de las que acabamos de repasar aquí– revelan una discrepancia entre lo que vemos (u oímos o tocamos) y lo que está ahí fuera, haríamos bien en no juzgar las experiencias perceptivas solo en función de su «precisión» en

cuanto a su coincidencia directa con la realidad. La percepción precisa –«verídica»– así entendida es una quimera. La evolución ha diseñado la alucinación controlada de nuestro mundo perceptivo para que potencie nuestras posibilidades de supervivencia, no para que sea una ventana transparente a una realidad externa (una ventana que, en cualquier caso, tampoco tendría ningún sentido desde un punto de vista conceptual).[20] En los capítulos siguientes, nos detendremos más a fondo en estas ideas, pero antes, merece la pena que salgamos al encuentro de un par de potenciales objeciones.

Según la primera de ellas, el enfoque de la percepción como alucinación controlada niega ciertos aspectos del mundo real que son innegables. «Si todo lo que experimentamos no es más que una especie de alucinación –podría reprochar alguien–, ¿por qué no te arrojas delante de un tren, a ver qué te pasa?».

Nada de lo que aquí digo debería entenderse como una negación de la existencia de cosas en el mundo, ya sean trenes a toda velocidad, gatos o tazas de café. El «control» en esa alucinación controlada tiene tanta importancia como la «alucinación». Describir la percepción de este modo no significa que todo valga, sino que la manera en que las cosas del mundo *aparecen* en la experiencia perceptiva es una construcción del cerebro.

Dicho esto, es útil distinguir entre lo que el filósofo ilustrado John Locke llamó las cualidades «primarias» y las «secundarias». Locke postuló que las cualidades primarias de un objeto son aquellas que existen con independencia del observador, como el hecho de que ocupe un espacio, sea sólido o se mueva. Un tren que se acerca a toda velocidad posee cualidades primarias de ese tipo en abundancia, y por eso arrojarse delante de él es muy mala idea, tanto si lo estás observando como si no, y sean cuales sean tus creencias a propósito de la naturaleza de la percepción. Las cualidades secundarias son aquellas cuya existencia depende de un observador. Son propiedades de los objetos productoras de sensaciones (o «ideas») en la mente, y no se

puede considerar que tengan una existencia independiente en el objeto. El color es un buen ejemplo de cualidad secundaria, pues la experiencia cromática depende de la interacción que una clase particular de aparato perceptivo mantiene con un objeto.[21]

Desde el enfoque de la alucinación controlada, tanto las cualidades primarias de los objetos como las secundarias pueden dar origen a experiencias perceptivas a través de un proceso activo y constructivo. En ninguno de esos casos, sin embargo, el contenido de la experiencia perceptual será idéntico al de la cualidad correspondiente del objeto.

La segunda objeción a ese enfoque tiene que ver con nuestra capacidad para percibir cosas nuevas. ¿Acaso no necesitaríamos siempre mejores conjeturas posibles –elaboradas de antemano– para todo lo que podamos percibir? ¿Y no significaría eso que nos encontraríamos atrapados en un mundo perceptivo de lo ya previsto? Imagínate que nunca has visto un gorila, ni en la vida real, ni por la tele, ni en una película, ni tan siquiera en un libro, y que, un día, te encuentras inesperadamente con uno paseando por la calle. Yo te garantizo que ese día verás un gorila, y que esa será una experiencia perceptiva nueva y probablemente bastante alarmante para ti. En un mundo donde todo está previsto, ¿cómo es posible que ocurra algo así?

La respuesta corta es que «ver un gorila» nunca es una experiencia perceptiva completamente novedosa. Los gorilas son animales con brazos, piernas y pelo, y tú y tus antepasados ya habréis visto otras criaturas que presentan algunos de esos rasgos, si no todos ellos. En general, los gorilas son objetos que presentan unos bordes definidos (por peludos que sean), que se mueven de una forma bastante previsible y que reflejan la luz del mismo modo que otros objetos de similar tamaño, color y textura. La experiencia novedosa de «ver un gorila» está fundamentada, pues, sobre la base de unas predicciones perceptivas que actúan en múltiples niveles diferentes de detalle, y que se han adquirido a lo largo de múltiples escalas temporales diferentes –desde predicciones acerca de la luminancia y los bordes, hasta predicciones sobre las caras

y las posturas– que, juntas, esculpen una nueva mejor conjetura perceptiva posible para que tú veas un gorila por primera vez.

La respuesta larga implica aprender más sobre el modo en que el cerebro realiza el tremendamente complejo ejercicio gimnástico neural implicado en la inferencia perceptiva, y eso nos disponemos a hacer precisamente en el capítulo siguiente.

Fig. 7. Esto es lo que es. Imagen de referencia para la figura 6.*

5. LA MAGIA DE LAS PROBABILIDADES[1]

El reverendo Thomas Bayes (1702-1761), pastor presbiteriano, filósofo y estadístico que vivió gran parte de su vida en Tunbridge Wells, en el sur de Inglaterra, jamás llegó a publicar en vida el teorema que inmortalizó su nombre. Su *Ensayo encaminado a resolver un problema en la teoría del azar* fue presentado en la Royal Society, en Londres, dos años después de su muerte, por su compañero clérigo-filósofo Richard Price, y gran parte del trabajo pesado de los cálculos y las ecuaciones correría a cargo del matemático francés Pierre-Simon Laplace unas décadas después. Pero el de Bayes sería finalmente el apellido que quedaría asociado para siempre a un modo de razonamiento conocido como «inferencia de la mejor explicación» cuyas ideas son fundamentales para que entendamos mejor cómo se construyen las percepciones a partir de las mejores conjeturas que elabora el cerebro.

La bayesiana es una lógica de las probabilidades. Para ser más concretos, consiste en elaborar inferencias óptimas –lo que, hasta este punto, hemos venido llamando «mejores conjeturas»– en condiciones de incertidumbre. «Inferencia», un término que nos hemos encontrado varias veces en este libro, solo quiere decir alcanzar conclusiones basándose en las pruebas empíricas y la razón. La inferencia bayesiana es un ejemplo de razonamiento *abductivo,* diferente del *deductivo* y del *inductivo.* Deducir significa llegar a conclusiones solo mediante el uso de la lógica: si Jim es mayor que Jane, y Jane es mayor que Joe, entonces Jim es mayor que Joe. Si las premisas son verdaderas y se siguen las reglas de la lógica, las inferencias deductivas serán con toda certeza

correctas. La inducción, por su parte, consiste en alcanzar conclusiones mediante un proceso de extrapolación a partir de una serie de observaciones empíricas: el sol ha salido por el este durante toda la historia de la que se tiene registro, por consiguiente, cabe inferir que siempre sale por el este. A diferencia de las inferencias deductivas, las inductivas sí pueden ser erróneas: una inferencia como «las primeras tres bolas que extraje del saquito eran verdes, por lo tanto, todas las bolas que hay en él son verdes» puede ser cierta o no serlo.

El razonamiento abductivo –como es, por ejemplo, el formalizado por medio de la inferencia bayesiana– consiste exclusivamente en hallar la mejor explicación posible a un conjunto de observaciones cuando estas observaciones son incompletas, inciertas o ambiguas. Como en el caso del razonamiento inductivo, el abductivo puede errar. Cuando de buscar la «mejor explicación» se trata, podemos entender el razonamiento abductivo como una especie de razonamiento hacia atrás, es decir, que va de los efectos observados a sus causas más probables, más que hacia delante, de las causas a sus efectos (como ocurre con la deducción y la inducción).

He aquí un ejemplo. Una mañana, miras por la ventana de tu dormitorio y ves que el césped está mojado.[2] ¿Ha llovido de noche? Quizá, pero también es posible que te olvidaras de apagar el aspersor de riego del jardín. De lo que se trata es de encontrar la mejor explicación (o hipótesis) posible a lo que ves: *dado* que el césped está mojado, ¿cuál es la probabilidad de (i) que haya llovido de noche o de (ii) que te hayas dejado el aspersor encendido? Dicho de otro modo, queremos inferir *la causa más probable de los datos observados.*

La inferencia bayesiana te dice cómo hacerlo. Proporciona una manera óptima de actualizar nuestras creencias acerca de algo a medida que llegan nuevos datos. El teorema de Bayes es una fórmula matemática para pasar de lo que ya sabemos (la probabilidad *a priori)* a lo que deberíamos creer a continuación (la probabilidad *a posteriori)* basándonos en lo que estamos conociendo ahora (la probabilidad

condicional). Las probabilidades *a priori,* condicionales y *a posteriori* suelen llamarse «creencias» bayesianas porque representan estados de conocimiento, más que estados del mundo. (Nótese que una creencia bayesiana no es necesariamente algo que yo creo como individuo. Para el caso que aquí nos ocupa, tanto sentido tiene decir que mi córtex visual «cree» que el objeto que tengo delante es una taza de café como decir que creo que Neil Armstrong pisó la Luna).

Las creencias *a priori* son las probabilidades de que algo sea correcto previas a la llegada de nuevos datos. Digamos que la probabilidad *a priori* de que llueva por la noche es muy baja (a lo mejor, vives en un sitio como Las Vegas). Por su parte, la probabilidad *a priori* de que te dejes el aspersor encendido dependerá de la frecuencia con la que lo usas y de lo olvidadizo que seas. Digamos que es también baja, pero no tanto como la probabilidad *a priori* de que llueva.

Las probabilidades condicionales, dicho así, en un sentido muy general, son lo contrario de las probabilidades (o creencias) *a posteriori*. Suponen la formalización de un razonamiento «hacia delante», de las causas a los efectos: *dado* que llueva por la noche o que se deje el aspersor encendido, ¿cuál es la probabilidad de que el césped esté mojado? También pueden variar (como las creencias *a priori),* pero, de momento, supongamos que existe la misma probabilidad de que el césped se quede mojado después de una lluvia nocturna que después de un riego accidental a esa misma hora.

El teorema de Bayes combina probabilidades *a priori* y condicionales para calcular probabilidades *a posteriori* para cada hipótesis. El teorema en sí es simple: la probabilidad *a posteriori* no es más que la probabilidad *a priori* multiplicada por la condicional, y dividida por una segunda probabilidad *a priori* (la «de los datos», que, en este caso, es la probabilidad *a priori* de que el césped esté mojado, y de la que no hace falta que nos preocupemos más aquí, porque siempre es la misma para cada hipótesis).[3]

Al ver el césped mojado por la mañana, un buen bayesiano debería

decantarse por la hipótesis con la mayor probabilidad *a posteriori,* pues esta será la explicación más probable de los datos. Puesto que, en nuestro ejemplo, la probabilidad *a priori* de lluvia nocturna es menor que la de riego accidental, la probabilidad *a posteriori* de que haya llovido también será más baja. Un buen bayesiano optará, pues, por la hipótesis del aspersor. Esa hipótesis es la mejor conjetura bayesiana sobre las causas de los datos observados: es la «inferencia de la mejor explicación».

Si esto parece poco más que un ejercicio de mero sentido común, es porque, en este ejemplo en particular, así lo es. No obstante, existen numerosas situaciones en las que la inferencia bayesiana se aleja de lo que el sentido común podría indicarnos de entrada. Por ejemplo, es fácil concluir erróneamente que tienes una enfermedad grave tras haber dado positivo en alguna prueba médica, porque hay una tendencia bastante común a sobreestimar la probabilidad *a priori* de padecer enfermedades raras. Un test puede tener una precisión del 99 %, pero el hecho es que, si la prevalencia poblacional de la enfermedad en cuestión es lo bastante baja, un resultado positivo en una prueba no hará más que incrementar muy ligeramente la probabilidad *a posteriori* de que estés enfermo de eso.[4]

Volvamos a la situación del césped mojado y llevémosla un poco más allá. Imagina que, tras inspeccionar la hierba de tu jardín, echas un vistazo a la de tu vecino y ves que también la suya está mojada. Esta es una información nueva y muy relevante. Las probabilidades condicionales de cada una de las dos hipótesis son ahora diferentes: en el caso de la del aspersor, solo tu césped debería estar mojado, pero en el caso de la de la lluvia, ambos jardines deberían estarlo. (Las probabilidades condicionales, recuerda, van de las causas supuestas a los datos observados). Como buen bayesiano que eres, actualizas tus probabilidades *a posteriori* y llegas a la conclusión de que la lluvia nocturna es ahora la mejor explicación de lo que has visto, y cambias de opinión.

Un elemento muy potente de la inferencia bayesiana es el hecho de que tiene en cuenta la *fiabilidad* de la información a la hora de actualizar las mejores conjeturas. La información que (se estima que) es fiable debe tener una influencia mayor en las creencias bayesianas que aquella otra que (se estima que) no lo es. Imagina que la ventana de tu dormitorio está sucia y que has perdido las gafas. Parece que el césped de tu vecino está mojado, pero tu vista está tan mal y la ventana está tan mugrienta que esa nueva información es muy poco fiable y lo sabes. En ese caso, aunque la hipótesis de la lluvia se vuelve ligeramente más probable cuando ves lo que crees ver al otro lado de la valla, la hipótesis del riego accidental puede seguir llevando la delantera.

En muchas situaciones, el proceso de actualización de las mejores conjeturas bayesianas a medida que llegan nuevos datos se produce una y otra vez y da lugar a un ciclo de inferencias sin final. Con cada iteración, la antigua hipótesis *a posteriori* se convierte en la nueva hipótesis *a priori,* que es entonces utilizada para interpretar la siguiente ronda de datos para formar una nueva hipótesis *a posteriori* –una nueva mejor conjetura–, y el ciclo se repite. Si tu césped aparece mojado dos mañanas seguidas, en el segundo día tu mejor conjetura acerca de la causa debería estar condicionada por la que era tu mejor conjetura en el primer día, y así sucesivamente, un día tras otro.

La inferencia bayesiana se ha aplicado muy provechosamente en toda clase de contextos, desde las diagnosis médicas hasta la búsqueda de submarinos nucleares desaparecidos, y no dejan de surgir nuevas aplicaciones continuamente.[5] Hasta el propio método científico puede concebirse como un proceso bayesiano en el que las hipótesis científicas se van actualizando conforme se incorporan nuevos datos procedentes de experimentos. Esta concepción de la ciencia difiere tanto de la idea de los «cambios de paradigma» sobre la que escribió Thomas Kuhn (según la cual, las teorías científicas terminan por desmoronarse por completo en cuanto se acumulan demasiadas pruebas que las contradigan) como de las tesis «falsacionistas» de Karl Popper (para

quien las hipótesis se erigen y se contrastan una a una, como globos que se echan al aire y se intentan desinflar a perdigonazos). De hecho, en este terreno de la filosofía de la ciencia, la perspectiva bayesiana tiene más en común con las tesis del filósofo húngaro Imre Lakatos, cuyo análisis se alejó de las teorías sobre el ideal de la ciencia para centrarse en los factores que hacen que los programas de investigación científica funcionen en la práctica.[6]

Como es lógico, una concepción bayesiana de la ciencia significa que las creencias *a priori* que los científicos tienen sobre la validez de sus teorías influirán en la medida en que dichas teorías sean actualizadas o cuestionadas por los datos nuevos. Por ejemplo, soy consciente de que tengo la intensa creencia *a priori* de que los cerebros son máquinas de predicción de tipo bayesiano. Esta fuerte creencia no solo influirá en cómo interpreto la evidencia experimental, sino que también determinará la clase de experimentos que realizo, que irán dirigidos a generar nueva evidencia relevante para mis creencias. A veces, incluso me pregunto cuánta evidencia empírica haría falta para darle la vuelta a mi creencia bayesiana de que el cerebro es esencialmente bayesiano.[7]
Regresemos a nuestro cerebro imaginado, ese que está en silencio y a oscuras en el interior de su cráneo, enfrentado al reto de descifrar lo que hay ahí fuera, en el mundo. Pues bien, a estas alturas, ya estamos en disposición de apreciar hasta qué punto ese reto representa una oportunidad ideal para invocar la inferencia bayesiana. Y es que, cuando se trata de elaborar sus mejores conjeturas acerca de las causas de las ruidosas y ambiguas señales sensoriales que recibe, el cerebro no hace otra cosa que seguir los principios del reverendo Thomas Bayes.

Las creencias perceptivas *a priori* pueden codificarse en múltiples niveles de abstracción y flexibilidad, que van desde las creencias *a priori* muy generales y relativamente fijas, como que «la luz viene de arriba», a otras más situacionales, como que «el objeto peludo que se me acerca es un gorila». Las probabilidades condicionales con las que trabaja el cerebro llevan codificados unos mapas de correspondencia entre causas

potenciales y señales sensoriales. Son los componentes de «razonamiento hacia delante» de la inferencia perceptiva, y como ocurre con las creencias *a priori,* pueden operar a múltiples escalas de tiempo y de espacio distintas. El cerebro combina continuamente esas creencias *a priori* y esas probabilidades condicionales con arreglo al teorema de Bayes, de tal manera que, a cada fracción de segundo, se está formando una nueva creencia bayesiana *a posteriori* (una «mejor conjetura» perceptiva). Y cada nueva creencia *a posteriori* sirve como una *a priori* para la siguiente ronda de *inputs* sensoriales continuamente cambiantes. La percepción es un proceso circular, no una instantánea estática.

También aquí desempeña un papel importante la fiabilidad de la información sensorial. A menos que te encuentres en un zoo, la probabilidad *a priori* de estar percibiendo «un gorila» será muy baja cuando de pronto atisbes algo vagamente oscuro y peludo a lo lejos. Como, sea lo que sea, está lejos, la fiabilidad estimada de ese *input* visual será baja también, por lo que, en ese primer momento, tu mejor conjetura perceptiva difícilmente será «un gorila». Pero a medida que el animal se te aproxima, las señales visuales se van volviendo más fiables e informativas, y las mejores conjeturas que elabora tu cerebro van repasando entonces una serie de opciones –un gran perro negro, un hombre disfrazado de gorila, un gorila de verdad– hasta que percibes el gorila con la suficiente seguridad, con un poco de suerte, todavía a tiempo de escapar de él.

La manera más sencilla de entender las creencias bayesianas –las probabilidades *a priori,* condicionales y *a posteriori*– es representándolas como cifras entre 0 (para indicar una probabilidad cero) y 1 (para indicar un 100 % de probabilidad). No obstante, para comprender mejor cómo influye la fiabilidad de las señales sensoriales en la inferencia perceptiva, y de hecho, para ver cómo es posible que el cerebro ponga en práctica el teorema de Bayes, necesitamos profundizar un poco más y pensar en términos de *distribuciones de probabilidad.*

En la gráfica inferior, se muestra un ejemplo de distribución de

probabilidad para una variable X. En matemáticas, una variable no es más que un símbolo que puede tomar diferentes valores. Una distribución de probabilidad de X describe la probabilidad de que el valor de X se sitúe dentro de un intervalo concreto. Como se muestra en la gráfica, se puede representar mediante una curva. La probabilidad de que X esté dentro de un intervalo determinado viene dada por el área del espacio situado bajo esa curva entre el límite superior y el inferior de dicho intervalo. En este ejemplo, la probabilidad de que X tome un valor entre dos y cuatro es mucho mayor que la probabilidad de que dicho valor esté situado entre cuatro y seis. Como ocurre en todas las distribuciones de probabilidad, el área total que suma todo el espacio que hay bajo la curva es exactamente uno. Esto es así, porque, si se consideran todos los resultados posibles a la vez, es inevitable que suceda alguno de ellos.[8]

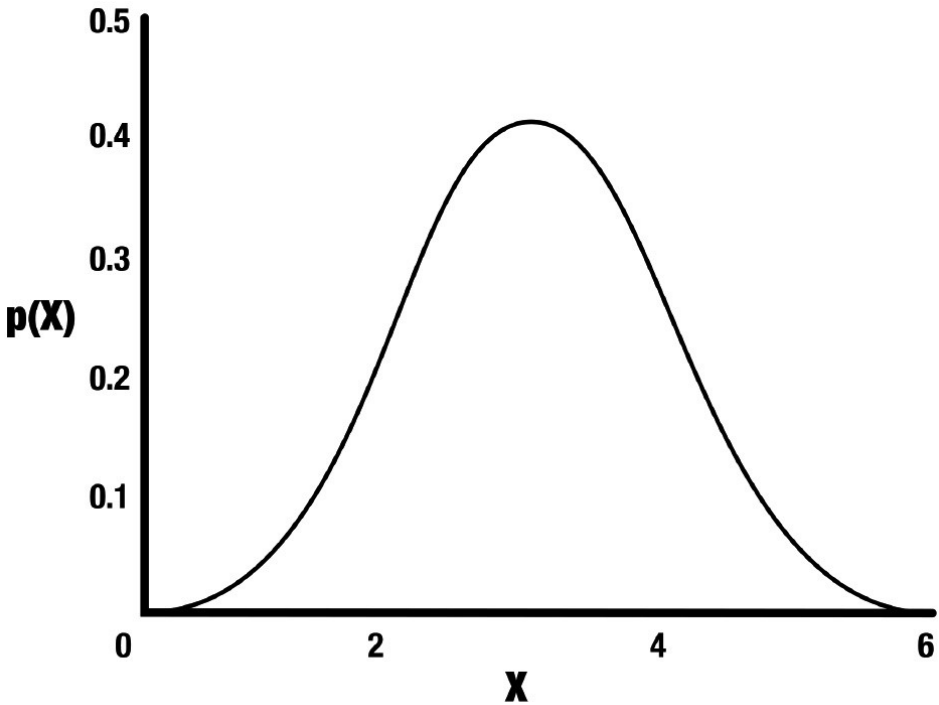

Fig. 8. Una distribución gaussiana de probabilidad

Las distribuciones de probabilidad pueden tener muchas formas diferentes. Una familia muy común de tales formas, de la que la curva aquí mostrada sería un ejemplo, es la de la llamada distribución «normal» o «gaussiana», también conocida como «campana de Gauss». Estas distribuciones están especificadas en su totalidad por un valor promedio o *media* (en el que la curva alcanza su pico, que, en este caso, es 3) y una *precisión* (lo extendida que está; cuanto mayor es la precisión, más concentrada es la distribución). Estas cantidades –media y precisión– son los *parámetros* de la distribución.[*]

Lo verdaderamente útil en este caso es que las creencias bayesianas se pueden representar como distribuciones de probabilidad gaussianas de este tipo. Intuitivamente vemos que la media especifica el contenido de la creencia, y la precisión expresa la confianza con la que el cerebro suscribe esa creencia. Una distribución con un pico más agudo (una precisión alta) se corresponderá con una creencia suscrita con un alto nivel de confianza. Como veremos, esta capacidad de representar la confianza –o fiabilidad– es la que dota a la inferencia bayesiana de su particular potencia.

Regresemos al ejemplo del gorila. Ya podemos concebir las creencias *a priori,* condicionales y *a posteriori* relevantes como distribuciones de probabilidad, cada una definida por una media y una precisión. Para cada distribución, la media señala la probabilidad de que sea «un gorila», y la precisión se corresponde con la confianza que el cerebro tiene en esa estimación probabilística.[9]

¿Qué ocurre cuando llegan nuevos datos sensoriales? El modo más fácil de ver el proceso de actualización bayesiana es representándolo gráficamente. En la gráfica siguiente, la curva de puntos representa la probabilidad *a priori* de encontrarse con un gorila. Esta curva tiene una media baja, que indica que los encuentros con gorilas son sucesos improbables, pero también tiene una precisión relativamente alta, que indica que esa creencia *a priori* se suscribe con una confianza elevada. La curva de rayas discontinuas es la probabilidad condicional, que se

corresponde con el *input* sensorial. En este caso, la media es más alta, pero la precisión, más baja: si de verdad hubiera un gorila ahí fuera, tal vez serían estos los datos sensoriales que recibirías, pero no estás demasiado seguro de que lo haya. La curva de raya continua es la creencia *a posteriori* de que haya ahí un gorila, dados los datos sensoriales de que dispones. Como siempre, estos resultados se obtienen aplicando la fórmula del teorema de Bayes. Cuando tratamos con distribuciones de probabilidad gaussianas, aplicar la fórmula de Bayes equivale a multiplicar la curva de puntos por la de rayas discontinuas, manteniendo el total del área bajo la curva resultante (la de la probabilidad *a posteriori)* limitado exactamente a 1.

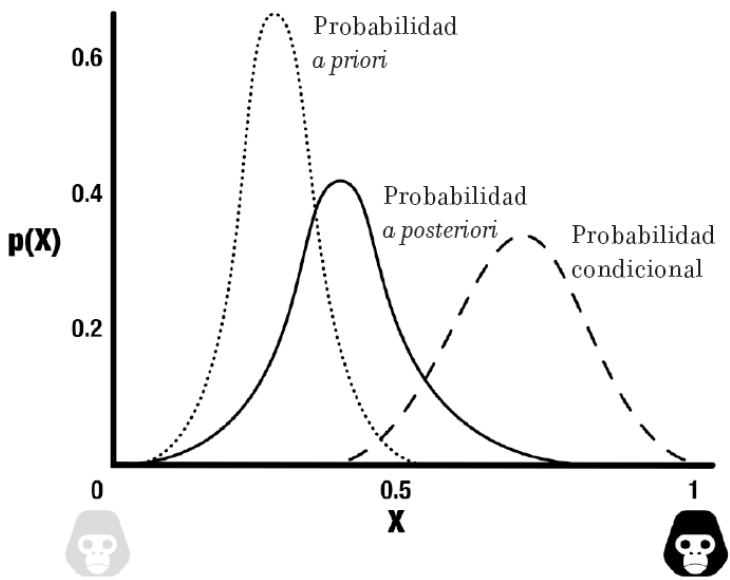

Fig. 9. Inferencia bayesiana con distribuciones gaussianas de probabilidad para elaborar mejores conjeturas posibles acerca del avistamiento de un gorila

Nótese que el pico de la distribución de probabilidad *a posteriori* se acerca más al de la de probabilidad *a priori* que al de la de probabilidad condicional. Esto se debe a que la combinación de dos distribuciones gaussianas depende tanto de las medias como de las precisiones. En este

caso, como la distribución de probabilidad condicional tiene una precisión relativamente baja –pues se juzgan poco fiables las señales sensoriales que indican la presencia de «un gorila»–, la de las mejores conjeturas *a posteriori* no se ha desplazado muy lejos de aquella. Sin embargo, en el instante siguiente en que mires, es posible que los datos sensoriales procedentes del gorila sean un poco más claros porque ya esté más cerca de ti; la nueva probabilidad *a priori* viene dada por la anterior probabilidad *a posteriori,* con lo que la nueva distribución de probabilidades *a posteriori* (la nueva mejor conjetura) se desplazará más cerca de la opción «un gorila». Y así seguirá ocurriendo hasta que sea ya hora de salir corriendo de ahí.

El teorema de Bayes proporciona un criterio de optimación para la inferencia perceptiva. Precisa el mejor de los casos en cada situación a propósito de aquello que un cerebro *debería* hacer a la hora de figurarse las causas más probables de los *inputs* sensoriales que recibe, ya sean estos gorilas, sillas rojas o tazas de café. Pero esto solo es parte de la historia. Lo que no hace el teorema de Bayes es especificar *cómo* consigue el cerebro, expresado en términos de mecanismos neuronales, tamañas gestas de optimación conjetural.

La respuesta a esta pregunta nos lleva de vuelta a la teoría de la percepción entendida como una alucinación controlada, y a su tesis central de que los contenidos conscientes están algo más que condicionados por las predicciones perceptivas: las propias predicciones *son* los contenidos.

En el capítulo anterior, introduje la idea de que la percepción se produce a través de un proceso continuo de *minimización del error de predicción.* Según esa idea, el cerebro está generando continuamente predicciones acerca de las señales sensoriales y comparando dichas predicciones con las señales sensoriales que le llegan a través de los ojos y los oídos…, y la nariz, y la piel, etcétera. Las diferencias entre las señales sensoriales predichas y las reales dan pie a errores de predicción. Las predicciones perceptivas fluyen predominantemente de arriba abajo (de dentro hacia

fuera), pero los errores de predicción fluyen de abajo arriba (de fuera hacia dentro). El cerebro usa esas señales de error de predicción para actualizar sus predicciones y estar listo para recibir la siguiente ronda de *inputs* sensoriales. Lo que percibimos viene dado por el contenido de todas las predicciones descendentes juntas, una vez minimizados (o «explicados») los errores de predicción en la mayor medida posible.[10]

El enfoque de la alucinación controlada comparte muchos elementos con otras teorías «predictivas» de la percepción y de la función cerebral, entre las que destaca la del *procesamiento predictivo*.[11] Existe, sin embargo, una importante diferencia de énfasis. El procesamiento predictivo es una teoría relativa a los *mecanismos* mediante los que el cerebro lleva a cabo la percepción (y la cognición y la acción). El enfoque de la alucinación controlada, sin embargo, se centra en cómo los mecanismos cerebrales explican las *propiedades fenomenológicas* de la percepción consciente. Dicho de otro modo, el procesamiento predictivo es una teoría del funcionamiento de los cerebros, mientras que el enfoque de la alucinación controlada toma esa teoría como base y la desarrolla para explicar la naturaleza de las experiencias conscientes. Ambas teorías están asentadas sobre un mismo e importante proceso fundamental: la minimización del error de predicción.

Y también es la minimización del error de predicción lo que conecta las alucinaciones controladas con la inferencia bayesiana, ya que toma una proposición bayesiana referida a lo que el cerebro *debería hacer* y la convierte en una proposición sobre lo que *hace* en realidad. Al minimizar errores de predicción en todo momento y lugar, lo que el cerebro está haciendo en el fondo es implementar el teorema de Bayes. Para ser más exactos, está usando una *aproximación* a la fórmula de Bayes. Esa conexión es la que permite que postulemos que el contenido perceptivo es una alucinación controlada que procede de arriba abajo, en vez de una «lectura» ascendente de los datos sensoriales.

Consideremos tres componentes centrales de la minimización del error de predicción en el cerebro: los modelos generativos, las jerarquías

perceptivas y la «ponderación de la precisión» de las señales sensoriales.

Los modelos generativos determinan el repertorio de cosas perceptibles. Para percibir un gorila, mi cerebro necesita estar equipado con un modelo generativo capaz de producir las señales sensoriales relevantes: es decir, aquellas que cabría esperar que percibiera si un gorila estuviera realmente en su presencia. Estos modelos proporcionan el flujo de predicciones perceptivas que se comparan con los datos sensoriales entrantes para formar errores de predicción que el cerebro trata de minimizar, con lo que, de paso, las propias predicciones se van actualizando.

Las predicciones perceptivas se desarrollan y se aplican para múltiples escalas de tiempo y de espacio, de tal forma que, al final, percibimos un mundo estructurado, lleno de objetos, personas y lugares. Una predicción de alto nivel de que lo que estamos percibiendo es un gorila da pie a otras predicciones de nivel inferior acerca de sus extremidades, sus ojos, sus orejas y su pelo, que, a su vez, generan a niveles más bajos aún una cascada de predicciones relativas a colores, texturas y bordes, y, ya en última instancia, otras referidas a variaciones en los niveles de brillo por todo el campo visual.[12] Estas jerarquías perceptivas actúan en todos los sentidos e incluso trascienden por completo los datos sensoriales. Si de pronto oigo la voz de mi madre, mi córtex visual puede ajustar sus predicciones e inducirme a suponer que esa figura que se acerca es mi madre. Si sé que estoy en el zoo, las regiones perceptivas de mi cerebro estarán más preparadas para ver un gorila que si voy caminando tranquilamente por la calle.

Merece la pena aclarar aquí que la «predicción» a la que nos referimos cuando hablamos de minimización del error de predicción no está necesariamente orientada al futuro. Simplemente significa ir más allá de los datos valiéndose de un modelo. En estadística, la esencia de la predicción reside en compensar la ausencia de datos suficientes. No importa si las predicciones así formadas se refieren al futuro –para el que es fácil que no dispongamos de «datos suficientes»– o a alguna

situación presente de la que se tiene un conocimiento incompleto.

El elemento clave final de la minimización del error de predicción es la ponderación de la precisión. Ya hemos visto que la fiabilidad relativa de las señales sensoriales determina la medida en la que se actualizan las inferencias perceptivas. Tu avistamiento inicial de un gorila lejano, o del césped de tu vecino a través de una ventana sucia, proporcionarán unas señales sensoriales de una baja fiabilidad, por lo que tus mejores conjeturas bayesianas no variarán por mucho. Ya hemos visto también que donde se capta el nivel de fiabilidad es en la precisión de la distribución de probabilidad correspondiente. Tal como nos mostraba la gráfica de la p. 105, los datos sensoriales con una precisión estimada baja tienen un efecto más débil sobre la actualización de las creencias existentes *a priori*.

Hablo de «precisión estimada», más que de «precisión» a secas, porque la precisión de las señales sensoriales no es algo que le venga directamente dado al cerebro perceptor. También tiene que inferirse. El cerebro se enfrenta no solo al desafío de averiguar las causas más probables de los *inputs* sensoriales que recibe, sino también al de calcular lo fiables que son los *inputs* sensoriales relevantes. Lo que esto quiere decir, en la práctica, es que el cerebro ajusta continuamente la influencia de las señales sensoriales en la inferencia perceptiva. Lo hace modificando provisionalmente la precisión estimada de las mismas. Eso es lo que quiere decir el concepto de «ponderación de la precisión».[13] Ponderar a la baja la precisión estimada significa hacer que las señales sensoriales tengan menor influencia a la hora de actualizar las mejores conjeturas, mientras que ponderarla al alza quiere decir lo contrario: que las señales sensoriales influyan con más fuerza en la inferencia perceptiva. Así pues, la ponderación de la precisión desempeña un papel esencial a la hora de coreografiar la delicada danza que necesitan mantener las predicciones y los errores de predicción para generar una mejor conjetura perceptiva.

Aunque parece complicado, lo cierto es que todos nosotros y nosotras

estamos íntimamente familiarizados con el papel de la ponderación de la precisión en la percepción. Incrementar la precisión estimada de las señales sensoriales no es otra cosa que «prestar atención». Cuando prestas atención a algo –por ejemplo, cuando te esfuerzas por ver si hay un gorila ahí, a lo lejos–, tu cerebro aumenta la ponderación de la precisión que atribuye a las señales sensoriales correspondientes, lo que equivale a incrementar su fiabilidad estimada (o a subir su «volumen»). Cuando entendemos la atención de este modo, podemos explicar por qué, a veces, no vemos cosas que están muy a la vista, incluso cuando las estamos mirando directamente. Si estás prestando atención a unos datos sensoriales –incrementando así su precisión estimada–, entonces otros pasarán a tener menos influencia en la actualización de tus mejores conjeturas perceptivas.

Precisamente llama la atención el hecho de que, en según qué situaciones, los datos sensoriales desatendidos puedan llegar a no tener influencia alguna. En 1999, el psicólogo Daniel Simons elaboró una conocida demostración en vídeo de ese fenómeno, que él denomina «ceguera por falta de atención».[14] Si no lo has visto todavía, te recomiendo que lo veas antes de seguir leyendo.[*]

He aquí lo que ocurre. En la demostración, los sujetos del experimento de Simons vieron un vídeo breve en el que aparecen dos equipos de tres personas cada uno. Las de un equipo van de negro y las del otro, de blanco. Cada equipo tiene un balón de baloncesto cuyos miembros se pasan entre sí, mientras se mueven siguiendo trayectorias aparentemente aleatorias. A los observadores se les pide que cuenten el número de pases que se han hecho entre sí solo los miembros del equipo blanco. Esto obliga a hacer un esfuerzo de concentración, ya que los seis jugadores se están moviendo de un lado a otro de la escena y se están pasando dos balones sin un criterio aparente.

Lo asombroso es que, mientras se están fijando en lo que allí ocurre, a la mayoría de personas que miran el vídeo les pasa completamente inadvertida una persona que, enfundada en un disfraz negro de gorila,

entra en escena por la izquierda (por la derecha de la pantalla), efectúa unos cuantos gestos gorilescos y sale de la escena por la derecha (por la izquierda de la pantalla). Si a los observadores se les muestra el mismo vídeo de nuevo, pero, esta vez, se les pide que se fijen en el gorila que allí sale, entonces lo ven al momento y, en muchos casos, se niegan a creer que están viendo el mismo vídeo. Lo que ocurre es que al centrar la atención en los jugadores de blanco, a las señales sensoriales de los jugadores de negro –y también del gorila– se les asigna una baja precisión estimada, por lo que tienen muy escasa (o nula) influencia en la actualización de las mejores conjeturas perceptivas.

Algo parecido me sucedió una tarde, hace muchos años, cuando iba en coche hacia mi lugar favorito para practicar surf en San Diego. Había girado a la izquierda justo donde acababan de instalar una señal de «prohibido girar a la izquierda», hacia una corta calle lateral que conduce a la orilla del océano en la barrio de Del Mar. Como no había ningún motivo obvio para poner aquella nueva señal, como otros vehículos delante del mío acababan de hacer ese mismo giro, como probablemente ya lo había hecho cientos de veces a lo largo de los años y como estaba muy cabreado por la injusta multa que me habían puesto, presenté una reclamación por escrito en la que declaré que la señal me había resultado literalmente invisible, por mucho que sí fuera visible «en principio». Mi defensa apelaba a los principios de la ceguera por falta de atención. Sí, allí había una señal nueva, pero debido al proceso de ponderación de la precisión en mi cerebro, no había podido percibirla. Llevé el caso incluso ante el tribunal estatal de tráfico de California, que, aun sin ser el Tribunal Supremo precisamente, sí es una instancia lo bastante alta como para que mi nombre y mi apellido aparecieran en el «orden del día de delitos» de aquella jornada. Incluso preparé una bonita presentación de PowerPoint para el juez, que no ayudó en lo más mínimo a mi causa.

También los magos sacan partido a la ceguera por falta de atención, aunque ellos posiblemente no describen su pericia en esos términos. En

particular, los trucos de magia en primer plano requieren de un magistral desvío de la atención de los espectadores para que estos no noten cuándo están colocando la reina de picas detrás de una oreja de la que, más tarde, parecerá que aparece… ¡como por arte de magia![15] Los carteristas expertos también se aprovechan de esa peculiaridad de la psicología perceptiva. Una vez fui testigo de cómo el maestro de carteristas Apollo Robbins les sisó sin esfuerzo alguno los relojes, las carteras y los monederos a algunos de mis colegas, gesta más asombrosa si cabe si se tiene en cuenta que muchos de ellos eran expertos en percepción, lo sabían todo sobre la ceguera por falta de atención y estaban perfectamente al tanto de lo que Robbins pretendía hacer.

Es tentador entender nuestra interacción con el mundo del modo siguiente. Primero, percibimos el mundo como es. Luego, decidimos qué hacer. Y después, lo hacemos. *Sentir, pensar, actuar.* Puede que las cosas parezcan así, pero, insisto, lo que las cosas parecen no es una buena guía para conocer cómo son en realidad. Es el momento de que la *acción* entre en escena.

La acción es inseparable de la percepción. La percepción y la acción están tan estrechamente entrelazadas que se determinan y se definen mutuamente. Cada acción altera la percepción porque cambia los datos sensoriales entrantes, y cada percepción es como es porque ayuda a guiar la acción. Sencillamente, la percepción carece de sentido si no hay acción. Percibimos el mundo que nos rodea para poder actuar eficazmente dentro de él, para alcanzar nuestros objetivos y –a largo plazo– favorecer nuestras posibilidades de supervivencia.[16] No percibimos el mundo como es, sino del modo que más útil nos resulta percibirlo.

Puede que la acción incluso preceda a la percepción. En vez de suponer que el cerebro elabora sus mejores conjeturas posibles para, a continuación, guiar la conducta con ellas, podemos imaginárnoslo más bien como un especialista en generar acciones que calibra continuamente valiéndose de señales sensoriales para cumplir de forma

óptima con los objetivos del organismo.[17] El cerebro se nos presenta así como un sistema intrínsecamente dinámico y activo que constantemente sondea su entorno y examina las consecuencias.[*]

En el procesamiento predictivo, la acción y la percepción son dos caras de una misma moneda. Ambas se sustentan en la minimización de los errores de predicción sensoriales. Hasta ahora, me he referido a ese proceso de minimización en términos de una actualización de predicciones perceptivas, pero esta no es la única posibilidad. Los errores de predicción también se pueden mitigar realizando acciones para cambiar los datos sensoriales y conseguir que los nuevos datos sensoriales encajen con una predicción ya existente. Minimizar el error de predicción por medio de la acción es lo que se conoce como *inferencia activa,* un término acuñado por el neurocientífico británico Karl Friston.[18]

Una manera útil de entender la inferencia activa es concibiéndola como una especie de predicción perceptiva que se cumple por sí misma: un proceso por el que el cerebro busca –mediante la realización de acciones– aquellos datos sensoriales que confirmen la validez de sus predicciones perceptivas. Puede tratarse de acciones tan simples como mover los ojos. Esta mañana, estaba buscando las llaves del coche entre el desorden habitual que hay encima de mi mesa. Con cada movimiento de mis ojos, mientras escrutaban veloces todos los rincones de aquel pequeño caos, no solo se actualizaban mis predicciones visuales a cada instante (taza vacía, taza vacía, clips sujetapapeles, taza vacía…), sino que mi foco visual iba interpelando continuamente a la escena que tenía ante mí hasta que la predicción perceptiva de las llaves del coche se cumplió.

Cualquier tipo de acción corporal hará que cambien de algún modo los datos sensoriales: mover los ojos, entrar en una habitación diferente, apretar los músculos del vientre… Incluso «acciones» de nivel superior, como solicitar un nuevo empleo o decidir que te casas, generarán una cascada de conjuntos de acciones físicas que modificarán los *inputs*

sensoriales. Todas las formas de acción tienen el potencial de suprimir errores de predicción sensoriales por medio de la inferencia activa, por lo que todas las formas de acción participan directamente en la percepción.

Como todos los demás aspectos del procesamiento predictivo, la inferencia activa depende también de la modelización generativa. En concreto, la inferencia activa descansa sobre la capacidad que tienen los modelos generativos de predecir las consecuencias sensoriales de las acciones. Se trata de predicciones del tipo «si miro ahí, ¿con qué datos sensoriales es probable que me encuentre?». Estas predicciones reciben el nombre de predicciones *condicionales,* porque se refieren a lo que *ocurriría* si algo fuese realmente así. Sin predicciones condicionales de esa clase, el cerebro no tendría modo alguno de saber qué acción –entre las incontables acciones posibles– sería la que más probablemente reduciría los errores de predicción sensoriales. Las acciones que mi cerebro predijo que serían las que, con mayor probabilidad, servirían para que encontrase las llaves del coche que no encontraba fueron del tipo «escruta visualmente tu mesa», no «mira atentamente por la ventana» ni «levanta los brazos y agítalos».

Además de para que se cumplan predicciones perceptivas ya existentes, la inferencia activa puede ser de ayuda también para mejorar esas predicciones. En escalas temporales breves, las acciones pueden recopilar nuevos datos sensoriales que ayuden a elaborar una mejor conjetura más depurada o a decidir entre hipótesis perceptivas alternativas. Al principio de este capítulo, ya vimos un ejemplo de cómo era posible discriminar mejor entre las hipótesis alternativas de la «lluvia nocturna» y el «aspersor que se quedó encendido por accidente» mirando el césped del vecino al otro lado de la valla. Otro ejemplo sería poner en orden todas las tazas que hay sobre mi mesa para ayudarme a mí mismo a encontrar las llaves de mi coche. En cada uno de esos casos, la elección de la acción relevante depende de que se disponga de un modelo generativo capaz de predecir cómo cambiarían los datos

sensoriales a consecuencia de esa acción.

A largo plazo, las acciones son fundamentales para el *aprendizaje,* que aquí significa mejorar los modelos generativos del cerebro a medida que se van revelando más cosas acerca de las causas de las señales sensoriales y la estructura causal del mundo en general. Cuando miro al otro lado de la valla para ayudarme a inferir las causas de que un césped en concreto esté mojado, también he aprendido más sobre lo que causa que el césped de los jardines se moje en general. En el mejor de los casos, la inferencia activa puede originar un círculo virtuoso en el que unas acciones bien seleccionadas descubren información útil acerca de la estructura del mundo, que es luego incorporada a unos modelos generativos mejorados, lo que, a su vez, puede permitir una inferencia perceptiva perfeccionada y puede orientar hacia la realización de nuevas acciones que, según las predicciones, proporcionen una información más útil todavía.[19]

El aspecto más contraintuitivo de la inferencia activa es esa idea de que la *acción misma* puede concebirse como una forma de predicción perceptiva que se cumple por sí misma. Desde esa perspectiva, las acciones no solo participan en la percepción, sino que *son* percepciones. Cuando muevo los ojos para buscar las llaves de mi coche, o mis manos para ordenar las tazas, lo que está ocurriendo es que las predicciones perceptivas acerca de la posición y el movimiento de mi cuerpo se están haciendo realidad por sí mismas.

En la inferencia activa, las acciones son predicciones *propioceptivas* que se cumplen por sí mismas. La propiocepción es una forma de percepción que mantiene al individuo al tanto de dónde está su cuerpo y cómo se está moviendo mientras registra señales sensoriales que fluyen desde receptores situados por todo el esqueleto y la musculatura. Probablemente no pensamos mucho en la propiocepción, porque, en cierto sentido, siempre está ahí, pero el simple hecho de que puedas tocarte la nariz con los ojos cerrados –¡prueba a hacerlo!– es una demostración del papel esencial que juega en todas nuestras acciones.

Desde la perspectiva de la inferencia activa, tocarme la nariz significa dejar que una sucesión de predicciones propioceptivas acerca del movimiento y la posición de la mano se cumplan por sí mismas y superen la evidencia sensorial que me dice que mis dedos en este momento *no* están tocando la nariz. La ponderación de la precisión vuelve a desempeñar aquí un importante papel. Para que las predicciones propioceptivas se hagan realidad a sí mismas, los errores de predicción que le están indicando al cerebro dónde está realmente el cuerpo deben atenuarse o «infraponderarse». Se podría entender como lo contrario de prestar atención, es decir, como una especie de «desatención» al cuerpo que le permite moverse.[20]

Cuando concebimos la acción de esta forma, se pone de relieve cómo esta y la percepción son dos caras de la misma moneda. En vez de que la percepción sea el *input* y la acción sea el *output* con respecto a una «mente» central, tanto la acción como la percepción son así formas de predicción cerebral. Ambas dependen de un proceso común de elaboración bayesiana de mejores conjeturas posibles, es decir, de una danza cuidadosamente coreografiada entre predicciones perceptivas y errores de predicción sensoriales, solo que con diferencias en quién lleva a quién.

Veamos una última vez cómo sigue nuestro cerebro imaginado, encerrado a cal y canto en su huesuda prisión. Ahora sabemos que no está ni mucho menos aislado. Nada en un torrente de señales sensoriales que le llegan del mundo y del cuerpo, y dirige continuamente acciones –predicciones propioceptivas que se cumplen por sí mismas– que esculpen de forma proactiva ese mismo flujo sensorial. La descarga sensorial entrante se encuentra con una cascada de predicciones descendentes, pero también hay señales de error de predicción que suben corriente arriba para estimular la elaboración de predicciones cada vez mejores y para provocar nuevas acciones. Este proceso circular origina una aproximación a la inferencia bayesiana, un «bayesianismo» de lo «suficientemente bueno» con el que el cerebro se conforma y se

reconforma a medida que evolucionan sus mejores conjeturas acerca de las causas de su entorno sensorial, y así se va creando un mundo perceptivo claro e intenso: una alucinación controlada.

Gracias a esta forma de entender las alucinaciones controladas, ahora tenemos buenas razones para admitir que las predicciones que proceden de arriba abajo hacen algo más que sesgar nuestra percepción: directamente, *son* lo que percibimos. Nuestro mundo perceptivo, con todos sus colores, formas y sonidos, no es nada más (ni nada menos) que la mejor conjetura que nuestro cerebro puede hacer sobre las causas ocultas de los *inputs* que le llegan desprovistos de color, forma y sonido.

Como veremos a continuación, no solo puede explicarse de ese modo el hecho de que experimentemos perceptivamente gatos, tazas de café y gorilas, sino *todos* los aspectos de nuestra experiencia perceptual.

6. LA PARTICIPACIÓN DEL ESPECTADOR

Nuestro viaje a la estructura profunda de la experiencia perceptiva comienza con una visita a la Viena de comienzos del siglo XX. Si hubieras podido estar en los cafés, los salones y los fumaderos de opio de tan elegante ciudad durante aquellos años, tal vez te habrías encontrado con unos cuantos personajes notables. Allí estaban los filósofos del Círculo de Viena, un grupo en el que se incluían Kurt Gödel, Rudolf Carnap y, de vez en cuando, Ludwig Wittgenstein. Allí estaban también los pioneros de la pintura expresionista Gustav Klimt, Oskar Kokoschka y Egon Schiele, así como el historiador del arte Alois Riegl. Y, por supuesto, también estaba Sigmund Freud.

En el fluido ambiente intelectual de la Viena de la época, las dos culturas del arte y de la ciencia se entremezclaron como pocas veces lo han hecho. Allí no se situaba a la ciencia por encima del arte, como tan habitualmente se hace cuando se considera que el arte y las reacciones humanas que suscita son cosas que precisan de explicación científica. Tampoco el arte se situaba a sí mismo más allá del alcance de la ciencia. Artistas y científicos –y sus respectivos críticos– eran aliados en sus intentos de comprender la experiencia humana en toda su riqueza y variedad. No es de extrañar que el neurocientífico Eric Kandel bautizase ese período precisamente con el nombre de «la edad de la comprensión» en su libro *La era del inconsciente.*[1]

Una de las ideas más influyentes nacidas en esa era de la comprensión fue la de la «participación del espectador», introducida inicialmente por Riegl y popularizada tiempo después por una de las mayores figuras de

la historiografía del arte en el siglo XX, Ernst Gombrich, nacido en Viena, por cierto, en 1909.[2] La idea de ambos subrayaba el papel desempeñado por el observador –o espectador– al «completar» con su imaginación una obra de arte. La participación del espectador es esa parte de la experiencia perceptiva que aporta el propio perceptor y que no se encuentra en la obra del arte (ni en el mundo) en sí.[3]

El concepto de la participación del espectador es muy conectable con teorías predictivas de la percepción como la de la alucinación controlada. Según Kandel, «la idea de que la percepción del espectador implica una inferencia "descendente" convenció a Gombrich de que no existe ninguna "mirada inocente"; es decir, toda percepción visual se basa en la clasificación de conceptos y en la interpretación de información visual. No se puede percibir lo que no se puede clasificar».[4]

A mí, la participación del espectador me resulta particularmente evidente cuando me hallo ante cuadros de artistas como Claude Monet, Paul Cézanne o Camille Pissarro. De pie ante algunas de sus obras maestras del impresionismo –como *Escarcha. El viejo camino a Ennery,* pintura de 1873 que actualmente se expone en el Museo de Orsay de París–, me siento atraído a un mundo distinto.[5] Una de las razones de que obras como esta adquieran ese poder reside en el espacio que dejan para que el sistema visual del observador lleve a cabo su labor interpretativa. En el cuadro de Pissarro, las «raspaduras de la paleta […] sobre un lienzo sucio» (por emplear la descripción que hizo el crítico Louis Leroy) evocan con fuerza la impresión perceptiva de un campo crudamente escarchado.[6]

Los paisajes impresionistas pretenden eliminar al artista del acto de la pintura en sí y recuperar el «ojo inocente» del que hablaba Gombrich[7] a base de trasladar al lienzo las variaciones de brillo que forman la materia prima de la inferencia perceptiva, en vez del resultado final de ese proceso. Para ello, el artista debe desarrollar y desplegar una sofisticada comprensión del modo en que se producen los aspectos subjetivos, fenomenológicos, de la visión. Cada obra puede entenderse

como un ejercicio de ingeniería inversa del sistema visual humano, yendo desde el *input* sensorial hasta la experiencia subjetiva coherente. Los cuadros se convierten en experimentos sobre percepción predictiva y sobre la naturaleza de las experiencias conscientes a las que esos procesos dan lugar.[8]

Cuadros como el de Pissarro son más que ecos o presagios de una ciencia de la percepción, y la participación del espectador nos ofrece algo más que una versión de la minimización del error de predicción desde la perspectiva de la historia del arte. Lo que Gombrich y compañía aportan es un reconocimiento profundo de la naturaleza fenomenológica, experiencial, de la percepción: un reconocimiento que puede pasarnos fácilmente inadvertido entre los entresijos de las creencias *a priori,* las probabilidades condicionales y los errores de predicción.

«Cuando decimos que los borrones y pinceladas del lienzo impresionista "cobran vida de repente", queremos decir que se nos ha conducido a proyectar un paisaje sobre esos toques de pigmento».[9] Ahí, Gombrich capta algo que es esencial en la percepción consciente, algo que es válido más allá del arte y se extiende a la naturaleza de la experiencia en general.[10] Cuando experimentamos el mundo como si estuviera «realmente ahí», no se está produciendo una revelación pasiva de una realidad objetiva, sino una proyección clara y presente: un intento del cerebro por proyectarse y contactar con el mundo.

De vuelta en el laboratorio, ya se han intentado desentrañar las formas en que las expectativas perceptivas sustentan la experiencia subjetiva comenzando con unos experimentos sencillos. Una predicción o hipótesis de trabajo en esos experimentos es que, en principio, percibiremos más rápido (y más fácil) aquellas cosas que esperamos que aquellas otras que no esperamos.[11] Pues bien, hace unos años, Yair Pinto –que entonces era un investigador posdoctoral que trabajaba conmigo, y que ahora es un profesor ayudante en la Universidad de Ámsterdam– se propuso contrastar esa hipótesis centrándose (como solemos hacer en

nuestro campo) en la experiencia visual.

Yair utilizó un sistema llamado de «supresión por proyecciones rápidas continuas» *(continuous flash suppression,* o CFS), en el que, a una misma persona, se le muestran a su ojo izquierdo imágenes diferentes de las que se le muestran a su ojo derecho. A uno de los ojos se le enseña una figura (en este caso, de una casa o de una cara), mientras que al otro se le muestran dibujos rápidamente cambiantes de rectángulos superpuestos. Cuando el cerebro intenta fusionar las dos imágenes en una sola escena, no consigue hacerlo y los rectángulos cambiantes tienden a imponerse, de tal modo que, al final, son estos lo que la persona termina por ver conscientemente. La percepción consciente de la figura queda suprimida por la «continua proyección rápida» de formas rectangulares.[12] Nuestro propio experimento (ilustrado más abajo) empleó una versión de ese método en la que el contraste de las formas rectangulares comenzaba siendo alto e iba disminuyendo conforme avanzaba el experimento, mientras que el de la figura comenzaba siendo bajo y se iba incrementando. Eso significaba que, al cabo de unos pocos segundos (a lo sumo), la figura –de una casa o de una cara– se volvía conscientemente visible.[13]

Para averiguar cómo afectaban las expectativas perceptuales a la percepción consciente, se las indujimos a los participantes enseñándoles la palabra «casa» o la palabra «cara» antes de cada ensayo experimental. Cabe señalar que esas expectativas solo eran parcialmente válidas, pues, cuando se les enseñaba inicialmente la palabra «cara», luego solo aparecía la imagen de un rostro en el 70 % de los ensayos, pero no en el otro 30 %, donde aparecía una casa. Lo mismo sucedía a la inversa, es decir, cuando se les inducía a esperar que apareciera una casa. Midiendo el tiempo que tardaba cada imagen en aparecer de entre las proyecciones rápidas en paralelo, pudimos determinar con qué velocidad llegaban las personas a ver una imagen concreta (de una casa o de una cara) de forma consciente, en comparación con cuando no tenían expectativa alguna de lo que iban a ver.

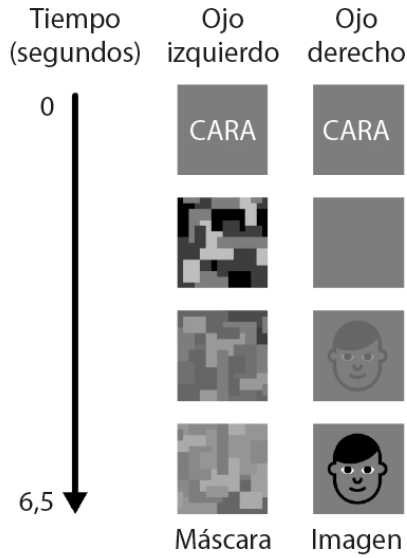

Fig. 10. Supresión por proyecciones rápidas continuas con contraste en aumento*

Como habíamos previsto, los participantes fueron más rápidos y más precisos a la hora de ver casas cuando eran casas lo que esperaban ver, y lo mismo podemos decir en el caso de las caras. La diferencia en velocidad fue pequeña (de aproximadamente una décima de segundo), pero fiable. En nuestro experimento, las expectativas perceptivas válidas condujeron realmente a unas percepciones conscientes más rápidas y precisas.

Nuestro estudio es uno de los que, en número creciente, se están dedicando a analizar qué ocurre cuando intervienen las expectativas perceptivas. En otro elegante experimento, Micha Heilbron, Floris de Lange y sus colegas del Instituto Donders de Nimega aprovecharon el llamado «efecto de superioridad de la palabra». Las letras sueltas –la «U», por ejemplo– son más fáciles de identificar cuando forman parte de una palabra como «CAUSA», que cuando aparecen en una sucesión no léxica de letras, como «AEUVR». El equipo de Lange mostró a unos voluntarios muchos ejemplos de palabras y de no palabras, siempre presentadas con fondos visualmente ruidosos. Y confirmaron el efecto de

superioridad de la palabra, porque, según vieron, las letras concretas eran más fáciles de leer si aparecían en una construcción léxica que si lo hacían en una que no correspondiera a vocabulario alguno.

El toque diferente en ese estudio vino del inteligente método que idearon los investigadores para estudiar la actividad cerebral de los participantes valiéndose de IRMf. Analizando los datos mediante una potente técnica llamada «lectura cerebral»,[14] hallaron que el marcador neuronal de una letra en el córtex visual era más «contrastado» –es decir, más distinguible de las representaciones neuronales de otras letras– cuando esa letra formaba parte de una palabra que cuando estaba inscrita en una no palabra. Esto significa que las expectativas perceptivas proporcionadas por el contexto léxico pueden modificar la actividad en las fases tempranas del procesamiento visual de un modo que potencie la percepción, tal como se postula desde el enfoque de la alucinación controlada.

Por muy reveladores que sean experimentos como estos, los entornos de laboratorio nunca reúnen la riqueza y la variedad de experiencias conscientes existentes fuera de ellos. Para salir del laboratorio y entrar en el mundo, es preciso que pensemos de un modo diferente.

Un día de verano de no hace mucho, y por primera vez en mi vida, me puse una minúscula cantidad de LSD bajo la lengua y me tumbé sobre la hierba a ver qué pasaba. El día era cálido, soplaba una brisa suave y el cielo era azul claro, salpicado de volutas nubosas. Al cabo de media hora, más o menos, como le había ocurrido a Albert Hofmann muchos años antes, el mundo comenzó a alterarse y mutar. Las colinas, el cielo, las nubes y el mar empezaron a latir, se volvieron más claros y presentes, deslumbrantes, como entrelazados y entretejidos con mi cuerpo, casi vivos en cierto modo. Como buen científico, traté de anotar lo que sucedía, pero cuando repasé las notas al día siguiente, me di cuenta de que mi esfuerzo había declinado casi al momento. Un recuerdo que ha permanecido conmigo es el de cómo las nubes adoptaron formas cambiantes pero definidas, y lo hicieron de un modo

que, al menos en parte, parecía estar bajo mi control. En cuanto una nube en particular empezaba por sí sola a parecerse a un caballo –o a un gato, o a una persona–, yo mismo me daba cuenta de que podía acentuar ese efecto sin apenas esfuerzo, y a veces hasta extremos absurdos. Hubo un momento en que toda una procesión de Cilla Blacks comenzó a desfilar sobre el horizonte.*

Nada mejor para desengañar a alguien que dude de que el cerebro sea el órgano de la experiencia que las alucinaciones inducidas por el LSD. Durante días, continué teniendo la impresión de que podía penetrar el «misterio» de mis experiencias perceptivas y experimentarlas (parcialmente, al menos) como las construcciones que son. Podía seguir experimentando ecos del medio paralelamente al mensaje perceptivo en sí.

Evidentemente, todos podemos adivinar caras en las nubes sin necesidad de un potenciador farmacológico (o, como mínimo, podemos ver lo que nos insinúan o sugieren esas sombras probabilísticas proyectadas en el cielo y sobre el cielo). El fenómeno general de ver figuras en las cosas recibe el nombre de *pareidolia* (del griego «junto a» e «imagen»). Para los seres humanos –y algunos otros animales–, la importancia de las caras implica que nuestros cerebros vengan preequipados con unas fuertes expectativas *a priori* relacionadas con los rostros. De ahí que todos, hasta cierto punto, tendamos a ver caras en las cosas, ya sean estas nubes, tostadas o incluso lavabos viejos (como el de la imagen de más abajo). Y como todos lo hacemos, no solemos considerar que la pareidolia sea una alucinación. La cosa es distinta cuando una persona esquizofrénica oye una voz que le ordena ejercer violencia contra sí misma, o que le dice que es Jesús renacido, y nadie más oye esa voz: entonces lo llamamos alucinación. También hablamos de una alucinación cuando, bajo los efectos del LSD, veo unas cuantas Cilla Blacks marchando por el cielo.

Fig. 11. Cuando vemos una cara en un lavabo[15]

Ahora sabemos que sería un error considerar cualquiera de esos fenómenos –por extraños que nos puedan parecer– como algo radicalmente distinto del proceso normal de generación de mejores conjeturas perceptivas. *Todas* nuestras experiencias, tanto si las calificamos de alucinatorias como si no, están fundadas en todo momento y lugar en una proyección de expectativas perceptivas en (y sobre) nuestro entorno sensorial. Lo que llamamos «alucinación» es lo que sucede cuando las creencias perceptuales *a priori* son inusualmente fuertes y anulan los datos sensoriales hasta el punto de que el cerebro comienza a perder toda comprensión de las causas mundanas de esos datos.

Inspirados por esa continuidad entre la percepción normal y la

alucinación, hemos estado explorando en nuestro laboratorio nuevas maneras de estudiar de qué modo las mejores conjeturas dan origen a la experiencia perceptiva, y nuestros experimentos nos han conducido hasta lugares ciertamente extraños.

Desde mi despacho, si subes dos tramos de escaleras y te abres paso entre las entrañas del viejo Departamento de Química, encontrarás uno de nuestros espacios de laboratorio improvisados, cuya ubicación y función están indicados en un papel pegado en la puerta con masilla: «VR/AR lab». Ahí utilizamos las tecnologías (en rápido desarrollo) de la realidad virtual y la realidad aumentada (RV y RA, o «VR/AR» en inglés) para investigar la percepción del mundo y del yo por vías imposibles de otro modo. Hace unos años, decidimos construir una «máquina de alucinaciones» para comprobar si podíamos generar experiencias alucinatorias de una forma controlable experimentalmente estimulando unas creencias perceptivas *a priori* de carácter hiperactivo.[16] El director del proyecto fue Keisuke Suzuki, investigador posdoctoral sénior de nuestro laboratorio, del que es el experto de cabecera en RV.

Lo primero que hicimos fue grabar con una cámara de vídeo de 360 grados unas imágenes panorámicas de un entorno del mundo real. Elegimos en concreto la plaza central del campus de la universidad un martes a la hora del almuerzo, cuando los estudiantes y el personal suelen darse una vuelta por el mercado semanal de comida que se instala allí ese día. Luego procesamos la grabación con un algoritmo diseñado por Keisuke basado en el de procesamiento de imágenes Deep Dream («sueño profundo») de Google, a fin de generar una alucinación simulada.

El algoritmo Deep Dream implica hacer que una red neuronal artificial a la que se ha entrenado para reconocer objetos dentro de las imágenes opere en sentido inverso. Las redes de ese tipo[17] están compuestas por múltiples capas de neuronas simuladas y sus conexiones están dispuestas de tal modo que se asemejan, en cierto modo, a la vía ascendente («de abajo arriba») en un sistema visual biológico. Dado que

estas redes solo contienen conexiones ascendentes, es fácil entrenarlas con métodos de aprendizaje automático. A la red concreta que usamos se la había entrenado para que identificara más de mil clases diferentes de objetos dentro de las imágenes, entre ellas, numerosas razas de perro. Y lo cierto es que realiza esa función de una manera excelente, llegando incluso a distinguir entre diferentes variedades de *husky* que a mí me parecen todas iguales.

La manera convencional en que se utilizan esas redes es mostrándoles una imagen y, a continuación, preguntándoles qué «piensan» que hay en ella. Sin embargo, el algoritmo Deep Dream le da la vuelta a ese procedimiento, pues le dice a la red que hay un objeto determinado presente y lo que esta actualiza entonces es la imagen.[18] Dicho de otro modo, el algoritmo está proyectando una predicción perceptiva sobre (y en) una imagen para darle un exceso de participación del espectador. En el caso de la máquina de alucinaciones, lo que hicimos fue aplicar ese procedimiento fotograma a fotograma a todo el vídeo panorámico, y añadimos algunas campanadas y silbidos para tratar la continuidad entre imágenes y cosas como esas. Volvimos a reproducir la película así procesada *(deep-dreamed)* en el visor de unas gafas de realidad virtual para que quienes se las pusieran pudieran mirar a su alrededor y experimentar esa RV de forma inmersiva, y así nació nuestra máquina de alucinaciones.

La primera vez que la probé, sentí una experiencia mucho más intensa de lo que me esperaba.[19] Aunque no fue, ni de lejos, como un viaje «a tope» de ácido ni como una alucinación psicótica (al menos, hasta donde yo conozco), doy fe de que la transformación del mundo que viví allí fue sustancial. Esta vez, no había Cilla Blacks, pero sí perros, y partes de perro, que surgían orgánicamente de todos los puntos de la escena a mi alrededor, y lo hacían de un modo que parecía completamente diferente de un simple «corta y pega» de imágenes de perro sobre una película preexistente (como muestra, véase un fotograma en la figura de esta página). La fuerza de esta máquina de alucinaciones reside en su

capacidad para simular los efectos de las mejores conjeturas (formadas desde arriba y hacia abajo) que nos dicen que allí hay perros presentes, pero haciéndolo de un modo que sintetiza –exagerándolo– el proceso por el que percibimos e interpretamos las escenas visuales en el mundo real.

Fig. 12. Fotograma de un vídeo de la «máquina de alucinaciones»

Programando la máquina de alucinaciones de formas ligeramente diferentes, podemos generar tipos distintos de experiencias alucinatorias. Por ejemplo, si retocamos la actividad en una de las capas intermedias de la red –en vez de en la capa del *output–,* terminamos teniendo alucinaciones de partes de objetos, no de objetos enteros. En este caso, la escena que se nos aparece se nos inunda de ojos, orejas y patas, un revoltijo de partes de perro que invade la totalidad de tu mundo visual. Y si retocamos capas aún más bajas, llegamos a lo que no hay mejor modo de describir que como alucinaciones «geométricas», en las que rasgos de bajo nivel del entorno visual –bordes, líneas, texturas,

patrones– adquieren una inusual nitidez y prominencia.

La máquina de alucinaciones es un ejercicio de lo que podríamos llamar «fenomenología informática», es decir, del uso de modelos informáticos para tender puentes explicativos entre los mecanismos y las propiedades de la experiencia perceptiva. Tiene un valor inmediato, que es el de casar la arquitectura informática de la percepción predictiva con la fenomenología de la alucinación. De este modo, podemos empezar a entender mejor por qué unos tipos determinados de alucinación son como son.[20] Pero más allá de esta aplicación, cabe buscar también ahí la que, para mí, es una conclusión más interesante, que es la de que arrojando este tipo de luz sobre las alucinaciones, igualmente podemos comprender mejor la experiencia perceptiva normal, cotidiana. La máquina de alucinaciones pone de manifiesto de un modo personal, inmediato y claro, que lo que llamamos alucinación es una forma de percepción incontrolada. Y que la percepción normal –en el aquí y el ahora– es, de hecho, una alucinación de carácter controlado.

De todos modos, sería una pena que el enfoque de la alucinación controlada se circunscribiera únicamente a explicar cosas como «veo una mesa porque esa es la mejor conjetura a la que ha llegado mi cerebro sobre las causas del actual *input* sensorial que está recibiendo». (O, para el caso, en vez de una mesa, una cara, un gato, un perro, una silla roja, un cuñado, un aguacate, una Cilla Black…). De hecho, yo creo sinceramente que podemos ir mucho más allá y que también nos sirve para explicar aquello que yo denomino la «estructura profunda» de la percepción: las maneras en las que el contenido consciente aparece en nuestra experiencia tanto en el tiempo como en el espacio, además de en diferentes modalidades.

Tomemos la observación, en apariencia trivial, de que nuestro mundo visual se compone en gran medida de objetos y de los espacios que median entre ellos. Cuando miro la taza de café sobre mi mesa, en cierto sentido percibo su parte posterior, aun cuando no sea la parte que estoy viendo directamente. La taza se me presenta como algo que ocupa un

volumen definido, a diferencia de una taza de café en una foto o en un dibujo. He ahí la fenomenología de «lo objetual». Lo objetual es una propiedad de cómo los contenidos conscientes visuales aparecen en general, más que una propiedad de una experiencia consciente puntual cualquiera.

Lo objetual es un rasgo dominante de la experiencia visual, pero no es universal. Si, en un día soleado, te fijas en una franja uniforme de cielo azul, no tendrás la impresión de que el cielo es «un objeto que está ahí». Y si miras directamente al sol y luego miras hacia otro lado, la imagen residual que habrá quedado grabada en la retina no será algo que experimentarás como un objeto, sino como un pequeño fallo temporal de la percepción. Pueden establecerse distinciones similares para otras modalidades perceptivas: las personas que padecen acúfenos no experimentan esos molestos sonidos como algo que se corresponda con cosas realmente existentes en el mundo. (De hecho, por eso se les llama a veces «zumbidos en los oídos»).

Hace tiempo que los artistas han constatado la relevancia de lo objetual para la percepción humana. La omnipresente *La traición de las imágenes* (véase la página 159), de René Magritte, explora la diferencia entre un objeto y la imagen de un objeto. Buena parte de la creación cubista de Pablo Picasso investiga cómo nuestra percepción de lo objetual depende de nuestra perspectiva en primera persona. Sus cuadros descomponen y recomponen objetos de múltiples maneras, y los representan desde diversas perspectivas al mismo tiempo. Podríamos considerar que estas pinturas (y otras parecidas) exploran los principios de lo objetual desde el enfoque de la participación del espectador.[21] La obra de Picasso en particular arrastra al observador hacia la creación imaginativa de objetos perceptivos entre un batiburrillo de posibilidades. Como dijo el filósofo Maurice Merleau-Ponty, el pintor investiga mediante su pintura los medios a través de los que un objeto se hace visible a nuestros ojos.[22]

En la ciencia cognitiva, donde más a fondo se ha examinado la

fenomenología de lo objetual ha sido en la «teoría de la contingencia sensomotora».[23] Según esta teoría, lo que experimentamos depende de cierto «dominio práctico» de cómo las acciones cambian los *inputs* sensoriales. Cuando percibimos algo, no es que las señales sensoriales transporten el contenido de lo percibido, sino que este emerge del conocimiento implícito que tiene el cerebro de cómo las acciones y las sensaciones se acoplan entre sí. Desde esa perspectiva, la visión –como todas nuestras modalidades perceptivas– es algo que un organismo *hace,* y no una información pasiva que se le suministra a una «mente» centralizada.

En el capítulo 4, describimos la experiencia de la rojez en términos de predicciones de base cerebral acerca de cómo las superficies reflejan la luz. Hagamos extensiva ahora esta explicación a lo objetual. Si sostengo un tomate frente a mí, percibo el tomate *como si tuviera una parte posterior,* lo cual no me ocurre en absoluto cuando miro la foto o el dibujo de un tomate (o de una pipa, como en el cuadro de Magritte), o un cielo azul despejado, ni cuando experimento una imagen residual en la retina. Según la teoría de la contingencia sensomotora, soy perceptualmente consciente de que el tomate tiene una parte posterior –pese a no verla directamente– en virtud de un conocimiento implícito, «cableado» en mi cerebro, acerca de cómo cambiarán mis señales sensoriales entrantes si le doy la vuelta a un tomate.

Fig. 13. René Magritte, *La traición de las imágenes* (1929)
© ADAGP, París, y DACS, Londres, 2021

El cableado necesario para ello tiene forma de modelo generativo. Como ya sabemos desde el capítulo anterior, los modelos generativos pueden predecir las consecuencias sensoriales de las acciones. Estas predicciones son «condicionales» o «contrafactuales» en el sentido de que tienen que ver con lo que *podría pasar* o lo que *podría haber pasado* con las señales sensoriales si se diera (o se hubiera dado) alguna acción en concreto. En un artículo científico que escribí hace unos años, postulé que la fenomenología de lo objetual depende de la *riqueza* de esas predicciones condicionales o contrafactuales.[24] Los modelos generativos que llevan codificadas múltiples y distintas predicciones de esa índole –como la de que un tomate tendrá la piel roja por todo alrededor– conducen a una fuerte fenomenología de lo objetual. Pero los modelos generativos que solo codifican unas pocas de esas predicciones (o ninguna de ellas), como los referidos a un cielo azul despejado o una

imagen residual en la retina, hacen que lo objetual sea débil o esté directamente ausente.

Otra situación en la que lo objetual suele estar ausente es la «sinestesia grafema-color».[25] El término sinestesia alude a una «mezcla de los sentidos». Las personas que tienen la variante de grafema-color tienen experiencias cromáticas cuando ven letras: por ejemplo, la letra «A» puede suscitar en ellas un rojo luminoso, con independencia del color real que tenga en aquella página y aquel momento. Aunque estas experiencias del color ocurren de forma sistemática y automática –es decir, que la persona siempre experimenta el mismo color cuando se encuentra con la misma letra en particular, y no necesita hacer esfuerzo consciente alguno para que eso sea así–, normalmente los sinéstetas no confunden sus colores sinestésicos con los colores «externos» o «del mundo real». Creo que esto se debe a que los colores sinestésicos, en comparación con los «reales», no se sustentan en un rico repertorio de predicciones sensomotoras. Un «rojo» sinestésico no varía apenas a medida que la persona se mueve, por ejemplo, o a medida que va cambiando la luz ambiental, por lo que ahí no puede haber ninguna fenomenología de lo objetual.

En nuestro laboratorio de RV, hemos empezado a poner a prueba algunas de estas ideas. Para un experimento reciente,[26] creamos una serie de objetos virtuales deliberadamente poco familiares, dibujados con bultos y salientes diferentes en cada caso, que los participantes veían a través de unas gafas de RV (véase la imagen de la página siguiente). Adaptamos el método de supresión por proyecciones rápidas utilizado en nuestro experimento anterior de la «cara-casa» de tal modo que, en esta ocasión, cada objeto fuese invisible inicialmente, pero terminase abriéndose paso progresivamente a la percepción consciente. Si en el experimento de la cara-casa, habíamos manipulado el hecho de que una imagen en particular fuese esperada o no, en esta otra adaptación con gafas de RV, pudimos manipular más bien la validez de las *predicciones sensomotoras* cambiando el modo en que los objetos

respondían a las acciones. Los participantes usaban un *joystick* para rotar los objetos virtuales, y nosotros hacíamos que estos reaccionaran como lo haría un objeto real o bien que se giraran de algún modo aleatorio e imprevisible. Predijimos que los objetos virtuales que seguían comportamientos normales alcanzarían la percepción consciente de los sujetos antes que aquellos otros que incumplían las predicciones sensomotoras, y eso precisamente fue lo que confirmaron nuestros resultados. Hay que reconocer que se trata de un experimento imperfecto porque usa la velocidad del acceso a la percepción consciente como valor representativo aproximado de la fenomenología de lo objetual. Pero, aun así, demuestra que la validez de las predicciones sensomotoras puede afectar a dicha percepción consciente de un modo específico y medible.

Fig. 14. Algunos de los objetos virtuales diseñados para parecer poco familiares

Entre las muchas ideas intuitivas pero falsas acerca de la percepción está la de que las variaciones en lo que percibimos se corresponden directamente con cambios acaecidos en el mundo. Pero el cambio, como lo objetual, es una manifestación más de la estructura profunda de la experiencia perceptiva. El cambio de la percepción no viene dado sin más por modificaciones concretas en los datos sensoriales. Percibimos cambios por los mismos principios de elaboración de mejores conjeturas que dan origen a todos los demás aspectos de la percepción.

Muchos experimentos han demostrado que el cambio físico –el cambio en el mundo– no es condición necesaria ni suficiente para que haya una

percepción de cambio. La imagen repleta de serpientes de más abajo nos proporciona un llamativo ejemplo de cómo, aunque nada se está moviendo en realidad, puedes tener una impresión perceptiva de movimiento, sobre todo si dejas que los ojos vayan recorriendo la imagen.[27] Lo que ocurre es que los pequeños detalles de la imagen, al ser vistos en la periferia de tu visión –con el rabillo del ojo–, convencen a tu córtex visual para que infiera movimiento aunque no haya ninguno.

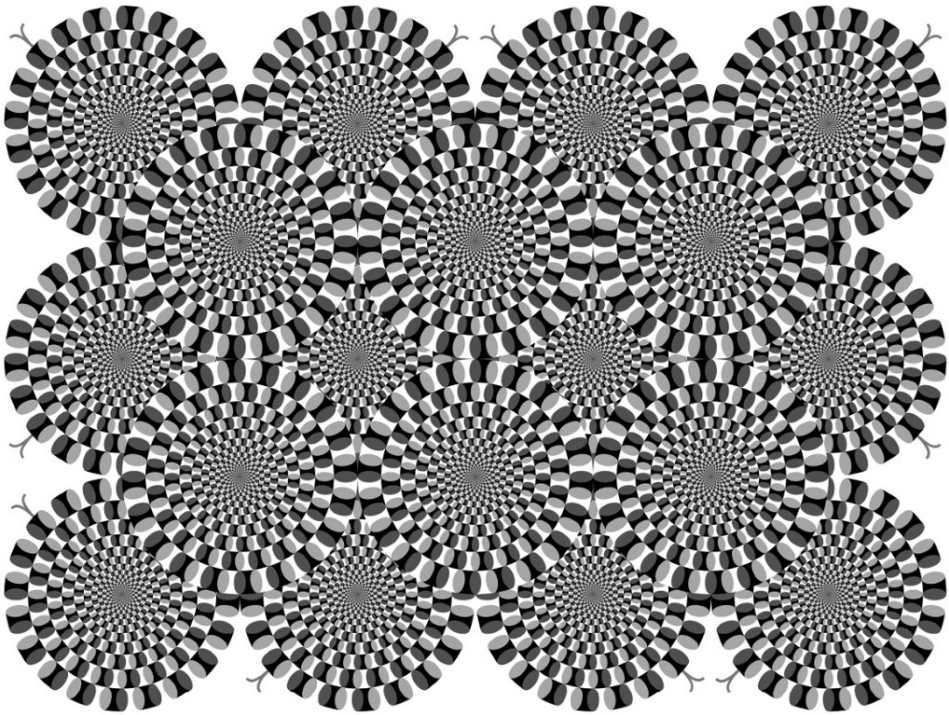

Fig. 15. Ilusión de las «serpientes giratorias»[*]
Cortesía de Akiyoshi Kitaoka

La situación contraria, la de que haya un cambio físico sin que se produzca una variación perceptiva, es la que se produce en la llamada «ceguera al cambio», que puede ocurrir cuando algunos aspectos de un entorno varían muy poco a poco, o cuando todo está cambiando a la vez y solo unos cuantos rasgos son relevantes. Hay un potente ejemplo en

vídeo de este fenómeno* en el que hasta la mitad de una imagen puede llegar a cambiar de color –de rojo a violeta– sin que la mayoría de personas noten el cambio para nada, aunque estén mirando directamente a la parte cambiante de la imagen. (Eso sí, solo funciona cuando no se las ha avisado de antemano de ese cambio de color. Si tratan activamente de detectarla, la variación es fácil de ver). Este ejemplo guarda cierta similitud con la ceguera por falta de atención que describí en el capítulo anterior (cuando los observadores no ven el gorila que aparece de improviso en medio de otras personas que están jugando a baloncesto). La diferencia, en este caso, es que lo que las personas no están acertando a ver es el «cambio» en sí.

Hay quienes piensan que la ceguera al cambio saca a relucir un dilema filosófico: cuando la imagen ya ha cambiado de color, ¿sigues experimentando el color rojo (pese a que ahora ya es violeta) o estás experimentando el violeta?, en cuyo caso, cabría preguntarse qué estabas experimentando antes, dado que no experimentaste cambio alguno. La solución consiste en negar la premisa de la pregunta y admitir que la *percepción de cambio* no es lo mismo que el *cambio de percepción.* La experiencia de cambio es una inferencia perceptiva más, otra variedad de alucinación controlada.

Y si las experiencias perceptivas del cambio son inferencias perceptuales, no lo son menos las experiencias perceptivas del tiempo.
El tiempo es uno de los temas más desconcertantes tanto en filosofía como en física, pero también en neurociencia. Los físicos se esfuerzan por entender qué es y por qué fluye (si es que fluye realmente), y el reto para los neurocientíficos no es menos espinoso. Todas nuestras experiencias perceptivas tienen lugar en el tiempo y a través del tiempo. Incluso nuestra experiencia del momento presente siempre parece desdibujarse entre un pasado relativamente fijado y un futuro parcialmente abierto. El tiempo fluye para nosotros también, aunque a veces parezca arrastrarse lentamente y, otras, dé la impresión de «volar».

Experimentamos los segundos, las horas, los meses y los años, pero no poseemos «sensores del tiempo» en el interior de nuestros cerebros. Para la visión sí contamos con fotorreceptores en la retina; para la audición, hay unas «células con cilios» en el oído interno; pero no hay ningún sistema sensorial específicamente dedicado al tiempo. Es más, aparte del ritmo circadiano (que es el que nos produce *jet lag,* entre otros fenómenos), no hay pruebas de que exista ningún «reloj neuronal» en el interior de la cabeza que mida nuestras experiencias en el tiempo (y que, en todo caso, sería un ejemplo privilegiado de lo que Daniel Dennett llamó una doble transducción, en la que una propiedad del mundo se recrea en el cerebro en beneficio de un supuesto observador interno). Pero, en vez de eso, las experiencias de tiempo –como las de cambio y como todas nuestras percepciones– son alucinaciones controladas también.[28]

Pero ¿qué las controla? Sin un canal sensorial específicamente dedicado a ello, ¿de dónde saldría el equivalente de los errores de predicción sensorial? Tenemos una solución simple y elegante a esa cuestión planteada por mi colega, el científico cognitivo Warrick Roseboom, que se incorporó a nuestro centro en 2015 y dirige ahora su propio grupo especializado en la percepción del tiempo. Su idea es que inferimos el tiempo basándonos no en el tictac de un reloj interno, sino en el ritmo de cambio de los contenidos de otras modalidades perceptivas. Y ha diseñado una manera muy inteligente de contrastar esa hipótesis.

Dirigido por Warrick, nuestro equipo creó una biblioteca de vídeos de diferentes duraciones y grabados en diferentes contextos: una calle urbana llena de gente, una oficina vacía, unas vacas pastando en un prado cerca de la universidad. Luego pedimos a unos voluntarios que miraran esos vídeos y estimaran cuánto duraba cada uno. Al hacerlo, todos evidenciaron unos sesgos característicos:[29] subestimaron la duración de los vídeos largos y sobreestimaron la de los cortos. También revelaron sesgos relacionados con el contexto de cada escena, pues

calcularon que las escenas más ajetreadas y llenas duraban más que las tranquilas, incluso en el caso de vídeos que tenían objetivamente la misma duración exacta.

Warrick mostró luego los mismos vídeos a una red neuronal artificial que mimetizaba el funcionamiento del sistema visual humano. Esta red era, en realidad, la misma que habíamos usado en nuestra máquina de alucinaciones. Para cada vídeo se calculó una estimación de la duración, basada –aproximadamente– en la tasa acumulada de cambio de actividad dentro de la red. En dichas estimaciones, no intervino «reloj interno» alguno. Pero, curiosamente, las estimaciones de la red neuronal y las humanas fueron prácticamente idénticas y evidenciaron los mismos sesgos con respecto a la duración y al contexto.[30] Esto nos muestra que la percepción del tiempo puede surgir –en principio, al menos– de una «mejor conjetura» a propósito del ritmo de cambio de las señales sensoriales, sin necesidad de un marcapasos interno.

Hace poco, hemos ampliado esa investigación buscando pruebas de la presencia de ese proceso en el cerebro. En un estudio dirigido por la investigadora posdoctoral Maxine Sherman, empleamos IRMf para registrar la actividad cerebral de unas personas mientras miraban un mismo conjunto de vídeos y estimaban la duración de cada uno de ellos.[31] Queríamos saber si podíamos usar la actividad en el córtex visual para predecir la duración aparente de cada vídeo como lo habíamos hecho en el estudio previo de Warrick, cuando usamos un modelo informático de la visión. Maxine descubrió que sí podíamos. La actividad cerebral en el sistema visual –aunque no en otras regiones del cerebro– era un buen «predictor» de la duración subjetiva. He ahí una prueba contundente de que las experiencias de duración surgen realmente de los procesos de formación de mejores conjeturas perceptivas, y no del tictac de ningún reloj neuronal.

Tampoco revelaron la existencia de un «reloj interno» otros experimentos que podían haberlo hecho. Mi favorito (con diferencia) es uno en el que se pidió a los sujetos que saltaran desde unas grúas.[32] El

neurocientífico David Eagleman se propuso testar la intuición común de que el tiempo subjetivo se vuelve más lento en momentos de mayor dramatismo, como los instantes previos a un accidente de tráfico. Su argumento era que esta ralentización subjetiva podía deberse a la existencia de un reloj interno que, en esos momentos, se acelera: cuantos más tictacs de dicho reloj durante un período temporal dado, más larga sería la duración percibida de ese período. Esto desembocaría, a su vez, en una «aceleración» del ritmo de percepción, ya que un reloj que corre más rápido implica una ampliación de la capacidad para percibir duraciones cortas.

Para testar esa idea, Eagleman y su equipo diseñaron un reloj digital especial que mostraba una serie de números que se sucedían tan rápido que era imposible leerlos en condiciones normales. Luego, convenció a unos cuantos voluntarios audaces para que efectuaran reiterados, sobrecogedores y adrenalínicos saltos al vacío mientras se miraban sus relojes y la vertiginosa sucesión de números que en ellos aparecía. La lógica del experimento era que, si hubieran tenido un reloj interno que se acelerara en esos momentos, los voluntarios deberían haber podido leer en esos instantes de caída libre los números que hasta entonces no habían sido capaces de distinguir. Pero no pudieron, por lo que el estudio de Eagleman no proporcionó prueba alguna de la existencia de un reloj interno. Evidentemente, la ausencia de pruebas no es una prueba de la ausencia, pero, aun así…, ¡menudo experimento!

Actualmente, a la vanguardia de nuestra investigación sobre la estructura profunda de la experiencia perceptiva, nos encontramos con un proyecto que estudia la percepción de la «realidad» en sí. Lo dirige Alberto Mariola, un talentoso estudiante de doctorado de nuestro laboratorio, y supone un novedoso diseño experimental de «realidad sustitutiva», como lo estamos llamando ahora mismo.[33] Por muy inmersivos que sean, los actuales entornos de RV siempre son distinguibles del mundo real para los sujetos. Los voluntarios que participan en nuestra máquina de alucinaciones siempre saben que lo

que están experimentando no es real, por muy psicodélicas que se vuelvan las cosas allí dentro. Pues bien, la realidad sustitutiva aspira a vencer esa limitación. El objetivo consiste en crear un sistema en el que las personas experimenten su entorno como si este fuese real, y que crean que es real, aunque de hecho no lo sea.

La idea es simple. Como para el caso de la máquina de alucinaciones, primero pregrabamos un vídeo panorámico, pero, esta vez, es una grabación del interior del mismo laboratorio de RV/RA exacto en el que llevamos a cabo nuestros experimentos. Cuando los voluntarios llegan al laboratorio, se sientan en un taburete en el centro de la estancia y se colocan un dispositivo de gafas de realidad virtual que lleva una cámara acoplada a su parte delantera. Se les invita entonces a que miren a su alrededor, en aquella sala, a través de la cámara montada en su dispositivo. En un determinado momento, cambiamos la fuente de las imágenes que les están llegando y hacemos que la cámara ya no les enseñe la escena del mundo real, sino el vídeo pregrabado. Curiosamente, la mayoría de personas, cuando ya están en esa situación, continúan experimentando lo que están viendo como si fuese «real», pese a que ya no lo es.

Con esta instalación experimental, ahora podemos testar ideas sobre las condiciones en las que las personas experimentan su entorno como si fuera real, y (lo que seguramente es más importante todavía) podemos comprobar qué es lo que hace que un aspecto tan presente –en condiciones normales– en nuestra experiencia consciente quede anulado de ese modo.[34] Estas situaciones pueden producirse y se producen no solo en casos como los de las imágenes residuales en la retina, sino también en trastornos psiquiátricos invalidantes, como los de despersonalización y desrealización, en los que puede haber una pérdida global de la realidad experimentada del mundo y del propio yo.

La investigación sobre lo que hace que la experiencia perceptiva parezca «real» nos lleva de vuelta a aquella idea de Wittgenstein acerca de la revolución copernicana: incluso cuando comprendemos cómo *son*

las cosas –que la Tierra gira alrededor del Sol o que la percepción es una alucinación controlada, por ejemplo–, en muchos sentidos las cosas continúan *pareciéndonos* como siempre nos ha parecido que eran. Cuando miro la silla roja del rincón de la habitación, me sigue *pareciendo* que tanto su rojez como su «sillidad» subjetiva son propiedades realmente existentes (verídicas) de una realidad independiente de la mente, y no elaboradas construcciones de un cerebro especialista en la formación de mejores conjeturas.

Hace mucho tiempo, en el siglo XVIII, David Hume escribió algo similar referido a la causalidad, otra característica muy presente en nuestra forma de experimentar el mundo. En vez de defender que la causalidad física era una propiedad objetiva del mundo, detectable por nuestros sentidos, Hume sostuvo que nosotros «proyectamos» la causalidad sobre el mundo basándonos para ello en nuestra percepción repetida de series de cosas que se suceden por estrecho margen temporal. Ni observamos ni podemos observar directamente «causalidad» alguna en el mundo. Sí, en el mundo ocurren cosas, pero lo que experimentamos como una causalidad es, en realidad, una inferencia perceptiva, de igual modo que todas nuestras percepciones son proyecciones de expectativas estructuradas que lanza nuestro cerebro sobre (y en) nuestro entorno sensorial: ejercicios de participación del espectador. Tal como lo explicó Hume, la mente tiene una gran propensión a extenderse hacia el mundo y desparramarse por él, de tal manera que «doramos y teñimos» los objetos naturales «con los colores tomados del sentimiento interior».[35] Y no solo con colores, sino con formas, olores, «sillidades», cambios, duraciones y causalidades, también: todas las características de primer y segundo plano de nuestros mundos perceptuales. Todas ellas son proyecciones humeanas, aspectos de una alucinación controlada.

¿Por qué experimentamos nuestras construcciones perceptivas como si fuesen objetivamente reales? Según el enfoque de la alucinación controlada, la finalidad de la percepción es guiar la acción y la

conducta: potenciar las posibilidades de supervivencia del organismo. Percibimos el mundo no *como es,* sino *como nos es más útil* percibirlo. Por lo tanto, tiene sentido que las propiedades fenomenológicas –como la rojez, la «sillidad», la «Cilla-Blackeza» o la causalidad– *parezcan* propiedades objetivas y verídicas de un entorno externo existente. Podemos reaccionar de un modo más rápido y eficaz a algo que ocurre en el mundo si lo percibimos como algo que existe realmente. La «eso-está-ahí-dad» inherente a nuestra experiencia perceptiva del mundo es, creo, una característica necesaria de un modelo generativo capaz de prever su flujo sensorial entrante con el propósito de guiar exitosamente el comportamiento.

Dicho de otro modo, aunque las propiedades perceptivas dependen de modelos generativos descendentes, nosotros no experimentamos esos modelos *como* tales modelos. Más bien, percibimos *con* y *mediante* nuestros modelos generativos, y cuando lo hacemos, por mero mecanismo, surge todo un mundo estructurado.[36]

Al comienzo de este libro, prometí que, si seguíamos el enfoque del problema real, iríamos desgranando también el problema difícil de por qué (y cómo) un mecanismo físico de algún tipo da origen a, se corresponde con, o es idéntico a la experiencia consciente. ¿Estamos realizando algún progreso?

Sí. Partiendo del principio según el cual el cerebro debe inferir las causas ocultas de sus *inputs* sensoriales, hemos llegado hasta una nueva comprensión de por qué y cómo nuestro universo interior está poblado por tazas de café, por colores y hasta por causalidad: cosas todas ellas que *parecen ser* propiedades de una realidad objetiva externa, aunque ese *parecer ser* es en sí una propiedad de la inferencia perceptiva. Y es precisamente la propiedad del «parecer que es real» la que añade leña al fuego a las típicas intuiciones del dualismo referido a la relación entre la experiencia consciente y el mundo físico, unas intuiciones que nos conducen, a su vez, hasta la idea del problema difícil. Es el hecho de que nuestras percepciones tengan el carácter fenomenológico de «ser reales»

lo que hace que resulte extraordinariamente difícil apreciar que, en realidad, las experiencias perceptivas no se corresponden necesariamente (ni, de hecho, nunca) de forma directa con cosas que tengan una existencia independiente de la mente. Una silla existe de manera independiente de la mente; la «sillidad» no.

Desde el momento en que nos damos cuenta de esto, se nos hace más fácil reconocer que el problema difícil no lo es tanto, y que puede que ni siquiera sea un problema propiamente dicho. Si invertimos los términos, el problema difícil de la conciencia solo parece difícil de verdad si interpretamos el contenido de nuestra experiencia perceptual como si fuera algo que realmente existe ahí fuera, en el mundo. (Que es precisamente lo que la fenomenología de la percepción consciente normal nos anima a hacer automáticamente).

Como ocurría con el estudio de la vida hace un siglo, la necesidad de encontrar un «algo especial» de la conciencia remite en la misma proporción en que aumenta nuestra capacidad para distinguir diferentes aspectos de la experiencia consciente y para explicarlos en términos de sus mecanismos subyacentes. Disolver el problema difícil no es lo mismo que resolverlo, ni tampoco significa refutarlo de forma definitiva, pero sí es la mejor vía para realizar progresos en este ámbito, una vía mucho mejor que la de venerar la conciencia como si fuera un misterio mágico o que la de desestimar esta como si no pasara de ser un «no problema» metafísicamente ilusorio. Y, además, nuestra misión cobra impulso cuando nos damos cuenta de que no solo nuestras experiencias del mundo son construcciones perceptivas.

Ha llegado la hora de que nos preguntemos quién –o qué– es el generador de toda esa percepción.

III. EL YO

7. DELIRIO

En el verano de 2014, mi madre entró en estado vegetativo mientras estaba ingresada en la unidad de emergencias quirúrgicas del hospital John Radcliffe de Oxford. Padecía una encefalopatía no diagnosticada con anterioridad, una enfermedad del cerebro. La causa nunca llegó a identificarse del todo; la habían ingresado por un cáncer intestinal y los problemas neurológicos surgieron de forma imprevista. Temiéndome lo peor, regresé al país a toda prisa desde un congreso al que estaba asistiendo en Brisbane. Ella se recuperó, lentamente, pero el recuerdo de cómo se fue disolviendo en cuestión de horas nunca me ha abandonado. Ella misma recuerda muy poco de aquello, lo cual no deja de ser una bendición.

Cuatro años después, vuelve a ser verano (el del mundial de fútbol de 2018) y estamos en medio de una inusitada ola de calor. Esta vez, no hay ningún estado vegetativo que lamentar, pero mi madre –que ya tiene ochenta y tres años– sufre lo que acabo de descubrir que se llama un «delirio hospitalario», un tipo distinto de fragmentación de su sentido del yo y del mundo que la rodea. Dos semanas antes, hubo que llevarla de urgencias al Radcliffe porque se le había reproducido un fuerte dolor intestinal. Dos días después del ingreso, mientras esperábamos a ver si el problema del intestino se resolvía sin necesidad de intervención quirúrgica, comenzó a desarrollar alucinaciones y delirios intensos, y entonces fui en coche hasta allí desde Brighton para estar con ella.

El delirio –una palabra que nace de las profundidades del siglo XVI– procede del latín *delirare,* «desviarse», «desquiciarse». El diccionario lo

define como el «trastorno de la mente en que hay intranquilidad, alucinaciones, hablar incoherente, etcétera». A diferencia de la demencia, que es una dolencia crónica y degenerativa, el delirio suele ser temporal. Crece y decrece, aunque puede durar semanas. La palabra evoca en mí la imagen de los frenopáticos de la era victoriana, así que me sorprende oír que todavía se use como diagnóstico en un hospital inglés del siglo XXI. Pero, reflexionando un poco más sobre ello, veo que no es tanto una sorpresa como un recordatorio más del largo camino que la medicina psiquiátrica tiene aún por recorrer.

El estado de mi madre se corresponde con bastante exactitud con esa definición del diccionario. Cuando me la encuentro en el pabellón del hospital, está sentada en la silla, encorvada, seria, desgreñada, con la mirada perdida. Me habla de las personas a las que ha visto trepando por las paredes, y no logra recordar dónde está ni por qué se encuentra ahí. Está perdiendo el contacto con la realidad y con quién es.

Lo peor llega un viernes. No se fía de nadie y está convencida de que quieren hacer con un ella un experimento muy cruel y trascendental, y de que nosotros –yo incluido, pues, en su paranoia, yo suelo figurar como el maestro de ceremonias– estamos induciendo deliberadamente esas alucinaciones en su cerebro por medio de los medicamentos que, con a saber qué finalidad oculta y maligna, le insistimos que tome. Ella, que es encantadora y afable en su estado normal, les grita airada a las enfermeras, exige irse del hospital, intenta escapar más de una vez y ordena a los médicos que echen de allí a su hijo-científico loco. Esta no es mi madre. Es una mujer que se parece a mi madre, pero que no es ella.

Entre los factores de riesgo que favorecen el delirio hospitalario están la epilepsia, las infecciones, la cirugía mayor (o alguna dolencia que la requiera), la fiebre, la deshidratación, haber comido o dormido poco, los efectos secundarios de alguna medicación y uno muy importante: la falta de familiaridad con el lugar. Mi madre los cumple todos. De hecho, por ser este el tipo de escenarios con el que estos pacientes se

encuentran tan poco familiarizados es por lo que este delirio en concreto recibe el nombre de «hospitalario».

Hay pocos lugares más proclives a crear una disociación con el mundo real que una unidad de emergencias quirúrgicas. Los pitidos constantes, las luces que parpadean, la ausencia de prácticamente cualquier señal del mundo exterior (una vista mínima a través de una ventana, si hay suerte): todo el mundo de la persona queda allí reducido a una cama, una silla y puede que un pasillo. Añádele unos compañeros de viaje que padecen diferentes grados de angustia y desaliño, y un desfile incesante de personal de enfermería y de médicos internos y especialistas, todos diferentes aunque también parecidos entre sí. Un día es igual que otro. El delirio, pese a ser una emergencia médica, no suele ser reconocido ni tratado como tal. Los pacientes ingresan con una enfermedad física y *ese* es el objetivo del tratamiento que reciben, y no ningún problema mental o cerebral que sobrevenga.

Hasta un tercio de los pacientes ancianos que ingresan en atención de urgencia en los hospitales desarrollan delirio hospitalario, y la proporción aumenta en el caso de los que reciben cirugía.[1] Aunque suele remitir con el tiempo, puede tener consecuencias graves a largo plazo, entre las que se incluyen una reducción de la capacidad cognitiva, un aumento de la probabilidad de muerte en los meses siguientes y un incremento del riesgo de posteriores episodios de delirio y demencia.[2] Me desplazo hasta la casa de pueblo donde vive habitualmente mi madre, la misma en la que me crie, para traerle de allí algunos objetos familiares con la esperanza de que la reorienten un poco: restos del naufragio a los que ella pueda aferrarse para no hundirse. Una fotografía enmarcada, sus gafas, una rebeca, un viejo león de peluche de mi lejana infancia.

Los delirios concretos dentro de ese estado de delirio general rara vez son casuales. Los que está teniendo mi madre siguen una lógica retorcida. Ella cree –o, según ella, *sabe*– que es víctima de un retorcido experimento en el que todo el mundo está implicado y del que yo soy

cómplice también. Y lo curioso es que no deja de ser cierto que yo *realizo* experimentos con personas, y que ahora, cuando estoy en el hospital, suelo adoptar una identidad extrañamente fluida, parte hijo y también parte médico, tratando de reconfortarla al tiempo que repaso notas médicas y hablo en voz baja con los especialistas y los internos acerca de encefalopatías paraneoplásicas y de otras cosas horribles. Al final, nuestro cerebro siempre está tratando de llegar a conclusiones, de formarse sus mejores conjeturas.

Algunas alteraciones de la conducta de mi madre son muy sutiles. Sus frases no son corrientes fluidas de palabras; más bien parece como si pronunciara cada vocablo por separado. «No encuentro las gafas, no sé dónde están». Este estado se mantiene hasta muchos días después de que su delirio primario haya remitido. También va como con altibajos. Esta noche he notado un paso atrás, otra regresión, y ha pinchado el globo de mis esperanzas de que pueda volver pronto a casa.

Incluso los contornos de mi propia vida se me empiezan a hacer irreales. Soy el único pariente cercano que le queda, así que me toca a mí estar ahí por ella. Las mañanas y las noches las paso en el hospital, y las tardes, si tengo suerte, me pongo al día con el trabajo atrasado y salgo a pasear y a bañarme en el río Támesis. Esas tardes, voy a Port Meadow, un gran espacio verde donde hay gansos, cisnes, vacas y hasta caballos salvajes. Un lugar inesperado, incluso entre las múltiples capas de mundos que coinciden en una ciudad como Oxford. La ola de calor se prolonga ya desde hace semanas, así que los campos, normalmente embarrados, hacen pensar ahora en la sabana africana. Esta mañana me persiguió un caballo y se me aceleró muy fuerte el corazón mientras cruzaba el puente sobre el ferrocarril, de vuelta a la ciudad.

Es el día catorce en el hospital para mi madre y el día doce para mí. Su momento de confusión aguda ya ha pasado, aunque sigue estando cambiada y su estado fluctúa. Pero ahora se queda asombrada cuando le digo que estaba convencida de que estábamos experimentando con ella, y que trató de que me echaran de allí. Le tomo la mano, le digo que

todo saldrá bien y espero que vuelva en sí misma por completo.

Pero ¿qué *es* ese «uno mismo», ese «yo»? ¿Es algo de lo que podamos alejarnos primero y a lo que podamos regresar después?

No, tampoco el yo es lo que podría parecer que es.

8. ESPÉRATE A TI MISMO

Tal vez te parezca que el yo –tu yo– es «quien» está percibiendo lo que hay. Pero la cosa no funciona así. El yo es una percepción más, otra alucinación controlada, aunque de un tipo muy especial. Desde tu sentido de la identidad personal –la de ser un científico o un hijo, por ejemplo– hasta tus experiencias de tener un cuerpo o de simplemente «ser» un cuerpo, los múltiples y variados elementos de la yoidad son mejores conjeturas bayesianas, diseñadas por la evolución para mantenerte vivo.

Comencemos nuestra exploración del yo con una rápida excursión al futuro. Dentro de un siglo, más o menos, ya se han inventado unos dispositivos de teletransporte capaces de crear réplicas exactas de cualquier ser humano. Funcionan como las máquinas de *Star Trek:* escanean una persona hasta el último detalle –teniendo en cuenta la disposición de todas y cada una de sus moléculas– y, usando la información recopilada por el escáner, fabrican una segunda versión de esa persona en un emplazamiento distante, por ejemplo en Marte.

Tras algunas reticencias iniciales, la gente se ha acostumbrado bastante rápido a esta tecnología, porque supone un medio de transporte muy eficiente. Sus usuarios incluso se han hecho a la idea de que el original deba evaporarse necesariamente de inmediato en cuanto está creada su réplica (un procedimiento que hubo que integrar en ese sistema para evitar que se creara una explosión de individuos idénticos). Desde el punto de vista de una viajera –llamémosla Eva–, eso no plantea problema práctico alguno. Tras recibir las debidas garantías del ente

operador del servicio, Eva simplemente siente que ha desaparecido del lugar X (Londres) y ha reaparecido en el lugar Y (Marte) en un instante.

Un día, surge una complicación. El módulo de evaporación en Londres sufre un fallo de funcionamiento y Eva –o, como mínimo, la Eva que está en Londres– tiene la sensación de que no ha ocurrido nada y que sigue en la instalación del teletransporte sin haberse movido. Un pequeño inconveniente, nada más. Tendrán que reiniciar la máquina y probar de nuevo, o tal vez dejarlo para el día siguiente. Pero, en ese momento, un técnico entra apesadumbrado en la sala con una pistola en la mano. Murmura algo del tipo: «No te preocupes, te hemos teletransportado a Marte de forma correcta, como siempre, lo que pasa es que las normas dicen que tenemos la obligación de… Y, perdona, pero mira, tú misma firmaste este formulario de consentimiento…». Levanta la pistola poco a poco, como a su pesar, y Eva tiene entonces una sensación que nunca antes había tenido, como si, tal vez, este invento del teletransporte no fuese tan directo y simple como pensaba.

El propósito de este experimento mental, que se conoce con el nombre de la «paradoja del teletransporte»,[1] es poner de relieve algunos de los sesgos que la mayoría de nosotros tenemos cuando pensamos en lo que significa *ser un yo*.

Dos son los problemas filosóficos planteados por la paradoja del teletransporte. El problema relativo a la conciencia en general es si podemos estar seguros de que la réplica tendrá experiencias conscientes, o si será más bien un equivalente perfectamente funcional, pero sin universo interior. No es un problema que me interese mucho. Si la réplica se crea con la suficiente minuciosidad de detalle –idéntica en todas y cada una de sus moléculas–, no hay razón para dudar de que sea consciente, y de que lo sea en exactamente el mismo sentido que el original. Si la réplica no es completamente idéntica, entonces volveremos a entrar en argumentos relativos a los diferentes tipos de zombi filosófico, y no hace falta que volvamos a repasar todo eso de nuevo.

El otro (y más interesante) problema es el de la identidad personal. ¿Es la Eva de Marte (llamémosla Eva2) *la misma persona* que Eva1 (la Eva que sigue en Londres)? Tal vez nos sintamos inclinados a responder que sí, que lo es: Eva2 va a sentirse exactamente igual que Eva1 si esta hubiera sido teletransportada realmente de Londres a Marte de forma instantánea. Lo que, al parecer, importa de verdad para este tipo de identidad personal es la continuidad psicológica, más que la física.* Pero, claro, si Eva1 no se ha evaporado, *¿cuál es la Eva de verdad?*

Yo creo que la respuesta correcta (por extraño que parezca) es que *ambas* son la Eva real.

Nosotros, intuitivamente, tratamos nuestras experiencias del yo de un modo distinto de como tratamos nuestras experiencias del mundo. Es como si nos costara más resistirnos a creer intuitivamente que la experiencia de *ser uno mismo* nos está revelando una propiedad auténtica de cómo son las cosas –en ese caso, *un yo real–,* más que un conjunto de percepciones. Una consecuencia intuitiva de presuponer la existencia de un yo real es que solo puede haber uno (un yo), y no dos ni tres ni más.

La idea de que el yo es, en cierto modo, indivisible, inmutable, trascendental, sui géneris, se forjó en la fragua del ideal cartesiano del alma inmaterial, y continúa teniendo una profunda relevancia psicológica, sobre todo en las sociedades occidentales. Pero también ha sido sometida repetidas veces al escrutinio escéptico de muchos filósofos y religiosos, así como, en época más reciente, al de los psiconautas psicodélicos, los médicos y los neurocientíficos.

Kant, en su *Crítica de la razón pura,* sostuvo que el concepto del yo como «simple sustancia» es erróneo, y Hume redujo el yo a un «haz» de percepciones.[2] En tiempos mucho más recientes, el filósofo alemán Thomas Metzinger escribió un libro brillantísimo, titulado *Being No One* («Ser nadie»), que es una potente deconstrucción del yo singular.[3] Hace mucho que los budistas defienden que no existe un yo permanente; de hecho, a través de la meditación, intentan alcanzar unos estados de la conciencia totalmente desvinculados de yo alguno. En América del Sur

(y cada vez más, en todo el mundo), las ceremonias con ayahuasca despojan a las personas de su sentido del yo mediante una embriagadora mezcla de ritual y dimetiltriptamina.

En neurología, Oliver Sacks –entre otros– ha escrito largo y tendido sobre las muchas formas en que el yo se descompone a raíz de ciertas enfermedades o daños cerebrales, mientras que los pacientes con «cerebro dividido» (de los que hablamos en el capítulo 3) hacen que nos planteemos la posibilidad de que un yo pueda convertirse en dos. El caso más curioso de todos es el de los gemelos craneópagos, que no solo están unidos físicamente, sino que también comparten parte de sus estructuras cerebrales. ¿Qué sentido puede tener un yo individual si resulta que un gemelo craneópago puede sentir que el otro se está bebiendo un zumo de naranja?[4]

Ser uno mismo no es tan simple como parece.

Estamos de vuelta en las instalaciones de teletransporte. Eva1 ha logrado zafarse de la maniobra asesina del técnico y está empezando a comprender su nueva situación, mientras que Eva2 sigue tan tranquila en Marte, al margen del drama que se está desarrollando en la tierra.

Aunque ambas Evas son objetiva y subjetivamente idénticas en el instante de la replicación, sus identidades ya han comenzado a divergir a partir de ese momento. Como sucede con los gemelos idénticos cuando emprenden sus propias vidas particulares, el proceso se complica inevitablemente con el tiempo. Incluso aunque Eva1 hubiera estado justo al lado de Eva2, se habrían ido produciendo pequeñas diferencias entre los *inputs* sensoriales de cada una de ellas que habrían desembocado también en diferencias sutiles de comportamiento y, así, casi sin darse cuenta, Eva1 y Eva2 estarían experimentando cosas dispares, grabando recuerdos diferenciados, convirtiéndose en personas distintas.

Estas complejidades de la identidad personal surgen de maneras diferentes en cada uno de nosotros. La identidad de mi madre cambió extraordinariamente durante su delirio, y aunque ya se ha recuperado

de aquello, hoy me parece –a mí, al menos– distinta de quien era antes y, al mismo tiempo, reconocible como la persona que era (lo que, en el fondo, vendría a ser el mismo caso que el de las dos Evas). La relación entre Eva1 y Eva2 podría incluso parecerse un poco a la relación que se puede establecer entre quien eres ahora y quien eras diez años atrás (o quien serás dentro de una década).

Cuando hablamos de quien tú eres, o de quien yo soy (el yo que, subjetiva y objetivamente, es «Anil Seth»), las cosas no son tan simples como parecen de entrada. Para empezar, el sentido de la identidad personal –ese «yo» que hay detrás de los ojos– solo es uno de los varios aspectos de cómo aparece en la conciencia el hecho de «ser un yo».

De hecho, la que sigue es mi manera preferida de subdividir los elementos de un yo humano.

Hay experiencias de *yoidad corporeizada* que están directamente relacionadas con el cuerpo. Entre ellas, se incluyen los sentimientos de identificación con ese objeto concreto que resulta ser tu cuerpo: sentimos cierta noción de propiedad de nuestro cuerpo como no la sentimos respecto a otros objetos del mundo. Las emociones y los estados de ánimo son también aspectos de la yoidad corporeizada, igual que los estados de activación y de alerta. Y subyacentes a esas experiencias, podemos encontrar ciertos sentimientos informes (aunque más profundos) referidos al simple hecho de ser un organismo vivo corpóreo –de *ser un cuerpo–,* sin una extensión espacial ni un contenido específico claramente definibles. Volveremos a esta capa básica fundamental de la yoidad más adelante. De momento, bastará con que la identifiquemos con el «sentimiento de estar vivo».

Más allá del cuerpo en sí, también tenemos la experiencia de percibir el mundo desde un punto de vista particular, es decir, de tener una perspectiva en primera persona: un punto subjetivo de origen de la experiencia perceptiva que, por lo general, parece residir dentro de la cabeza, alojado en algún lugar entre los ojos y un poco por detrás de la frente. Nada ilustra mejor este *yo perspectivo* que el autorretrato que

pintó el físico austríaco Ernst Mach y que se conoce también por el título de «Vista desde el ojo izquierdo».

Las experiencias de *volición,* de que uno (o una) tenga *intención* de hacer cosas y sea la causa de cosas que pasan *(agencia),* son también centrales para la yoidad. Ese es el *yo volitivo.* Cuando se habla del «libre albedrío», se suele estar haciendo referencia a estos aspectos de la yoidad. Para muchas personas, la idea del «libre albedrío» capta ese aspecto de ser-un-yo que menos dispuestas están a entregar al estudio de la ciencia.

Todos estos modos de ser-un-yo pueden existir en un individuo ya desde antes de que este tenga concepto alguno de una identidad personal, que es esa identidad que se puede asociar a un nombre, una historia y un futuro. Como vimos con la paradoja del teletransporte, para que exista la identidad personal, tiene que haber una historia previa personalizada, un hilo de recuerdos autobiográficos, un pasado recordado y un futuro proyectado.

A este sentido de identidad personal, cuando surge, se le puede llamar el *yo narrativo.* Con su aparición, llega también la capacidad de experimentar emociones sofisticadas como el arrepentimiento, que es algo que va más allá de la mera desilusión. (Los seres humanos también podemos sufrir un «arrepentimiento anticipatorio», que es la sensación que tengo cuando estoy seguro de que lo que estoy a punto de hacer saldrá mal, y de que, a pesar de saber esto, lo haré de todos modos y tanto yo como otras personas sufrirán por ello). En ejemplos así, podemos ver las ramificaciones e interacciones que se producen entre los diferentes niveles de yoidad: el surgimiento de la identidad personal cambia el abanico de estados emocionales disponibles (incrementándolo) y, al mismo tiempo, aquella es definida en parte por estos.

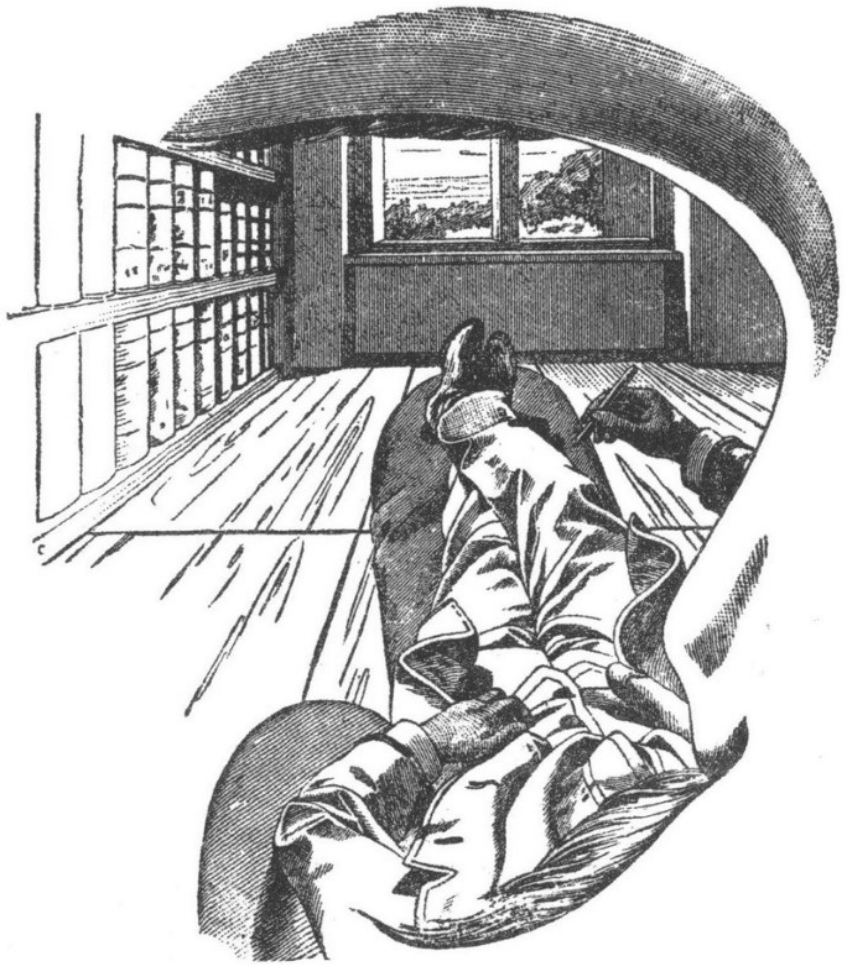

Fig. 16. Ernst Mach, *Autorretrato* (1886)

El *yo social* es el relativo a cómo percibo a otros que me perciben a mí. Es esa parte de mí que nace del hecho de que esté integrado en una red social. El yo social surge progresivamente durante la infancia y continúa evolucionando a lo largo de la vida, aunque puede desarrollarse de un modo distinto bajo trastornos como el del autismo.

La yoidad social trae consigo su propia gama de posibilidades emocionales, desde nuevas formas de sentirse mal –como la culpa y la vergüenza– hasta ciertas maneras características de sentirse bien, como el orgullo, el amor o el sentimiento de pertenencia.

Para cada uno de nosotros y nosotras, en circunstancias normales, estos elementos diversos de la yoidad están unidos en una sola pieza, subsumidos todos dentro de una experiencia unificada global: la experiencia de *ser uno mismo*. El carácter unificado de esa experiencia puede parecer tan natural –tanto como la unión perceptiva de color y forma cuando observas una silla roja– que es fácil darlo por sentado.

Pero nos equivocaríamos si lo hiciéramos. Del mismo modo que experimentar la rojez de algunas cosas no es una indicación de que exista un «rojo» externo en el mundo, las experiencias de una yoidad unificada no significan la existencia de un «yo real». De hecho, la experiencia de ser un yo unificado puede venirse abajo con facilidad.[5] El sentido de identidad personal, erigido sobre el yo narrativo, puede desgastarse o desaparecer por completo en situaciones de demencia o en casos graves de amnesia, y puede retorcerse y distorsionarse en casos de delirio (hospitalario o no). El yo volitivo puede estropearse en trastornos como la esquizofrenia o en síndromes como el de la mano ajena, en los que las personas experimentan una disminución de su sentido de conexión con sus propias acciones, pero también en el mutismo acinético, un trastorno por el que las personas dejan de mantener interacción alguna con su entorno. Las experiencias extracorporales y otros trastornos disociativos afectan al yo perspectivo, mientras que los trastornos relacionados con el sentido de propiedad corporal van desde el síndrome del miembro fantasma (experimentar sensaciones persistentes y, a menudo, dolorosas localizadas en una extremidad que ya no está ahí) hasta la somatoparafrenia (experimentar uno de tus miembros como si fuera de otra persona). Cuando padecen xenomelia (una variante extrema de la somatoparafrenia), las personas experimentan un deseo intenso de amputarse un brazo o una pierna, una

especie de remedio drástico a su sensación que, en alguna que otra rara ocasión, llevan realmente a la práctica.

El yo no es un ente inmutable que espía desde detrás de las ventanas de los ojos, desde donde observa el mundo de fuera y controla el cuerpo como un piloto controla un avión.[6] La experiencia de *ser yo,* o de *ser tú,* es una percepción en sí misma, o mejor dicho, un conjunto de percepciones, un haz fuertemente atado de predicciones codificadas a nivel neural y orientadas a mantener tu cuerpo con vida. Y yo estoy convencido de que esto es todo lo que necesitamos ser para ser quienes somos.

Pongamos el caso de la experiencia de identificarse con ese particular objeto-en-el-mundo que es tu cuerpo. Lo mudables que son esas experiencias y la precariedad con la que se reúnen son aspectos que no solo se hacen evidentes en trastornos como la somatoparafrenia o el síndrome del miembro fantasma, sino que también pueden revelarse por medio de experimentos de laboratorio simples. El ejemplo más famoso es el de la «ilusión de la mano de goma», que, tras haber sido descrito por primera vez hace más de veinte años, constituye actualmente una verdadera piedra angular de la investigación sobre corporeidad.[7]

La ilusión de la mano de goma es algo fácil de comprobar por ti mismo: lo único que necesitas es una persona voluntaria, unas cartulinas con las que formar una barrera, un par de pinceles y una mano de goma. En la imagen de la página siguiente, se muestra un esquema de cómo se coloca la otra persona participante. Esta voluntaria sitúa su mano (la real) a un lado de la división de cartulina, oculta a su vista. Al mismo tiempo, se le pone delante la mano de goma, justo en el sitio que habría ocupado su mano real normalmente y con la misma orientación. Entonces, el experimentador acaricia suavemente con los pinceles, moviéndolos hacia delante y hacia atrás, la mano real de la voluntaria y también la mano de goma. La idea es que, si se acarician ambas manos a la vez con los pinceles, de forma sincronizada, la voluntaria termina teniendo la extraña sensación de que la mano de goma forma parte de

su cuerpo, pese a que ella sabe que no es así; pero si se acarician una y otra mano de manera asíncrona, desacompasada, esa ilusión no llega a formarse y la persona no termina por asimilar nunca la mano de goma dentro de su experiencia de aquello que siente como su cuerpo.

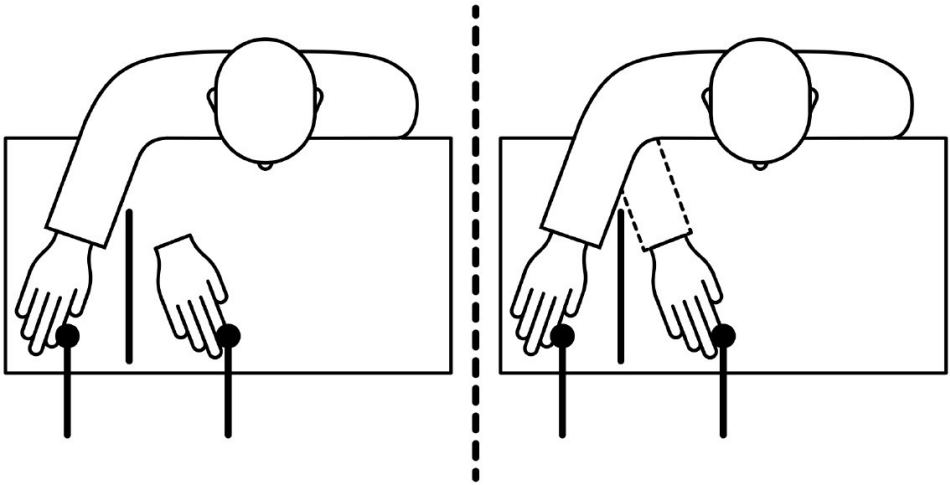

Fig. 17. La ilusión de la mano de goma. Cuando se acarician la mano de goma y la real simultáneamente (izquierda), la experiencia de propiedad corporal puede cambiar hasta el punto de que la persona llegue a sentir que la mano de goma forma parte de su cuerpo (derecha).

Esa descripción del experimento se ajusta muy bien a lo que les ocurre a algunas personas cuando se someten a él; desde luego, la sensación de que una mano a todas luces falsa forme parte de-algún-modo-aunque-no-del-todo de tu propio cuerpo es sin duda peculiar. Aun así, la experiencia real varía considerablemente de una persona a otra. Una manera de investigar esa variación consiste en amenazar de pronto la integridad de la mano de goma con un martillo o un cuchillo: en los casos en que la ilusión esté funcionando de verdad, la reacción en ese momento será mucho más intensa.

La ilusión de la mano de goma encaja muy bien con la idea de que las experiencias de propiedad corporal son formas especiales de alucinación controlada. En la situación en que ambas manos son acariciadas a la vez,

podría decirse que la combinación de *ver* como tocan la mano de goma con el pincel y, simultáneamente, *sentir* (que no ver) la caricia en la mano real le proporciona al cerebro suficiente evidencia sensorial como para crearse, como mejor conjetura perceptiva posible, la percepción de que la mano de goma forma parte del cuerpo. Y esto sucede en la situación sincronizada, pero no en la asíncrona, porque hay una expectativa previa que indica que las señales sensoriales que llegan simultáneamente al cerebro probablemente tienen una fuente común, que, en este caso, sería la mano de goma.

No son solo partes del cuerpo las que pueden experimentarse de forma diferente. También el cuerpo entero –y el origen de la perspectiva en primera persona– puede verse afectado.

En 2007, aparecieron casi a la vez en la prestigiosa revista *Science* dos artículos en los que se describía cómo podían usarse nuevos métodos de realidad virtual para generar una experiencia «extracorporal». Ambos experimentos se basaban en la ilusión de la mano de goma, aunque ampliada al conjunto del cuerpo. En uno de los estudios, llevado a cabo por un grupo de Lausana dirigido por Olaf Blanke, se pidió a unos voluntarios que se pusieran unas gafas a través de las que veían una representación en realidad virtual de la espalda de su propio cuerpo desde unos dos metros de distancia (véase la imagen de la página siguiente).[8] Desde esa perspectiva, observaban a continuación cómo un pincel acariciaba su cuerpo virtual; también en este caso las caricias podían ser sincrónicas o asíncronas con otras que se les aplicaban directamente sobre su propio cuerpo (real). Cuando las caricias estaban sincronizadas, la mayoría de participantes informaban haber sentido que su cuerpo virtual era, hasta cierto punto, su «propio» cuerpo; cuando se les pedía que anduvieran hasta donde sentían que estaba su cuerpo, su movimiento trazaba una deriva en el espacio tendente a la posición del cuerpo virtual.

Igual que la ilusión de la mano de goma insinúa que la sensación de propiedad corporal puede ser bastante flexible y momentánea,

experimentos como este otro (conocidos como «ilusiones del cuerpo entero») dan a entender que la propiedad subjetiva de todo el cuerpo y la ubicación de la perspectiva en primera persona también pueden manipularse sobre la marcha. Estos experimentos nos aportan pruebas fascinantes, además, de que las experiencias de «qué es mi cuerpo» pueden disociarse –al menos hasta cierto punto– de las experiencias de «dónde estoy».

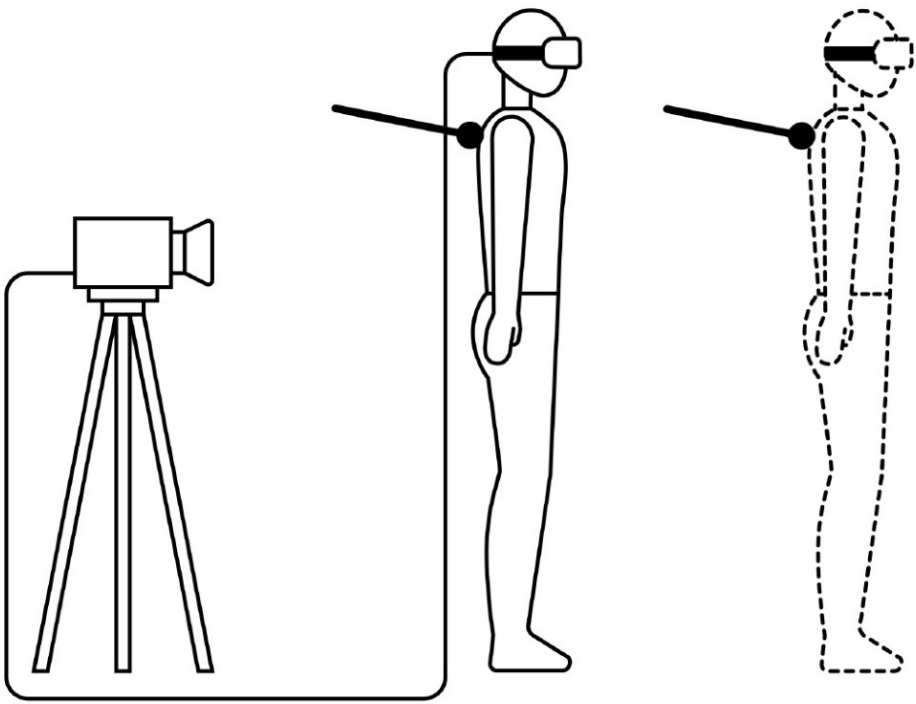

Fig. 18. Creación de una «ilusión del cuerpo entero»

Esta idea de que la perspectiva en primera persona de un individuo puede salirse de su propio cuerpo físico en forma de experiencias extracorporales (EEC) está muy arraigada en la historia y la cultura.[9] Los testimonios de EEC o de experiencias parecidas durante episodios traumáticos cercanos a la muerte en quirófanos, o en la periferia de

ciertos ataques epilépticos, han alimentado la fe en la existencia de una esencia inmaterial del yo. Después de todo, si puedes verte a ti mismo desde fuera, ¿no significa eso que la base de tu conciencia debe de ser separable de tu cerebro?

Sin embargo, si entendemos la perspectiva en primera persona como un tipo más de inferencia perceptiva, desaparece la necesidad de buscar anclajes milagrosos para un enfoque dualista. La idea de la «primera persona» como inferencia perceptiva se sustenta no solo en experimentos de realidad virtual como los de Olaf Blanke y otros investigadores, sino también en estudios de estimulación cerebral como los que se vienen desarrollando desde que el neurólogo canadiense Wilder Penfield realizara una serie de experimentos revolucionarios en la década de 1940.

Entre los pacientes de Penfield, había una mujer, conocida por las iniciales G. A., que, cuando se le estimulaba eléctricamente el giro temporal superior derecho (una parte del lóbulo temporal del cerebro), comenzaba a exclamar de manera espontánea: «Tengo la extraña sensación de no estar aquí… Es como si estuviera a medias aquí y a medias no».[10] El propio Blanke comenzó a fascinarse por las EEC cuando un paciente suyo, al serle estimulada una parte similar del cerebro –el giro angular, en la unión de los lóbulos temporal y parietal– informaba de una experiencia similar: «Me veo a mí mismo tumbado en la cama, desde arriba, pero solo me veo las piernas».[11]

El factor común a casos como esos es una actividad inusual en regiones del cerebro encargadas del *input* vestibular (el sistema vestibular es el que gestiona el sentido del equilibrio) que participan también en la integración multisensorial.[12] Parece ser que, cuando se altera la actividad normal en esos sistemas, el cerebro puede elaborar una «mejor conjetura» fuera de lo común acerca de la ubicación de su perspectiva en primera persona, aun cuando otros aspectos de la yoidad se estén manteniendo sin cambios.

También las experiencias similares a las EEC que a veces acompañan a

los ataques epilépticos pueden tener su origen en alteraciones de esos procesos. Estas experiencias se suelen dividir entre las *alucinaciones autoscópicas* –aquellas en las que ves tu entorno desde una perspectiva diferente– y las *alucinaciones heautoscópicas* (también conocidas como alucinaciones de *doppelgänger,* o de doble), en las que te ves a ti mismo desde una perspectiva distinta. Los testimonios documentados de experiencias de ambos tipos (que se remontan a cientos de años atrás) constituyen una corroboración adicional de la maleabilidad de la perspectiva en primera persona.[13] *

Cuando las personas informan haber tenido experiencias aparentemente sobrenaturales o, en cualquier caso, muy raras, como las EEC, debemos tomarnos en serio sus testimonios. Es probable que estén teniendo las experiencias que dicen tener. Hace milenios que las personas tienen experiencias extracorporales reales, lo que no significa que un yo inmaterial o un alma inmutable haya abandonado realmente su cuerpo físico en ninguno de esos momentos. Lo que esos testimonios revelan es que las perspectivas en primera persona se configuran por vías más complejas, provisionales y precarias que cualquiera de aquellas a las que jamás podamos tener acceso subjetivo directo alguno.

En el mundo virtual, la capacidad de alterar perspectivas en primera persona está generando algunas aplicaciones fascinantes, muchas de ellas alimentadas por un fenómeno que lleva el intrigante nombre de ilusión de «intercambio de cuerpos», descrito en un estudio de 2008 dirigido por el investigador sueco Henrik Ehrsson.[14] El experimento del intercambio de cuerpos está diseñado de tal modo que dos personas llevan puestas sendas gafas de realidad virtual, cada una con una cámara incorporada. Intercambiando las señales de las cámaras respectivas, cada persona puede verse desde el punto de vista de la otra. El efecto solo empieza a notarse de verdad cuando ambos participantes se dan la mano. Se supone que el hecho de ver y, al mismo tiempo, sentir el apretón de manos proporcionará una estimulación multisensorial que, combinada con expectativas generadas «desde

arriba», hará que cada persona sienta como si, de pronto, hubiera pasado a estar situada en el cuerpo de la otra y se estuviera dando la mano a sí misma. Puede decirse, pues, que esta experiencia te pone –aunque sea en un sentido virtual– en la piel de otra persona.

Yo mismo probé personalmente la sensación del intercambio de cuerpos en un pequeño encuentro en el que participé en Ojai (California) en el invierno de 2018. Estaba allí con Daanish Masood, un negociador para la paz de Naciones Unidas que también es investigador en el campo de la realidad virtual. Masood llevaba años colaborando estrechamente con BeAnotherLab, una creación de Mel Slater, neurocientífico instalado en Barcelona. El objetivo de BeAnotherLab es adaptar la tecnología de intercambio de cuerpos a unos novedosos dispositivos de «generación de empatía».[15] Para ello parten de la hipótesis según la cual, cuando una persona experimenta percibir el mundo desde dentro del cuerpo virtual de otro individuo, se genera en ella de manera natural una empatía hacia la situación del otro.

Daanish se había traído a su equipo hasta Ojai para hacer una demostración de su sistema, llamado The Machine to Be Another («La máquina para ser otro»). Su mecanismo añade al principio básico del intercambio de cuerpos un ingenioso elemento coreográfico que potencia aún más el efecto. Dos participantes se colocan las gafas de realidad virtual y, para empezar, se miran el regazo y ven el cuerpo de la otra persona en vez del suyo propio. A partir de ahí, van realizando una serie de movimientos coordinados conforme a unas instrucciones detalladas, y obedeciendo estas al dedillo, cada persona comienza a tener la sensación de que su nuevo cuerpo responde a las órdenes que ella le envía, lo que refuerza su experiencia de ser el otro. Al cabo de un rato, se les ponen delante unos espejos y cada participante ve la imagen reflejada del otro, pero como si fuera él mismo o ella misma. Finalmente, se descorre la cortina que separaba a las dos personas, y cada una empieza entonces a mirarse a sí misma desde dentro del cuerpo de la otra. Acto seguido, se acercan la una a la otra y se dan un

abrazo.

Cuando me llegó el turno de probarlo, me tocó intercambiar mi perspectiva propia con la de una mujer de condición social bastante elevada y unos setenta y pico años de edad. La experiencia fue inesperadamente intensa. Recuerdo que miré hacia abajo, flexioné mi (su) mano y noté –con cierta sorpresa– las rutilantes zapatillas deportivas que yo (ella) llevaba puestas. Las del espejo y el abrazo final fueron sensaciones particularmente potentes, aunque no estaba muy seguro de si era por la experiencia de sentir que habitaba el cuerpo de otra persona, o por la de verme desde la perspectiva de otro individuo. Fue más tarde, a la hora de la cena, cuando se me ocurrió preguntarme por lo extraño que debió de haber sido para mi compañera de experimento, también, haberse visto transportada de repente a la perspectiva en primera persona de un neurocientífico inglés mestizo y con calzado anodino.

Me resulta fascinante que se puedan manipular con semejante facilidad ciertos aspectos tan familiares y asumidos de nuestra yoidad (como la propiedad corporal subjetiva y la perspectiva en primera persona), ya sea mediante manos falsas y pinceles, ya sea mediante las nuevas tecnologías de la realidad virtual y aumentada. No obstante, hay un límite a lo lejos que esa clase de manipulaciones pueden llegar. Como ya he comentado, la experiencia típica con la ilusión de la mano de goma es sentir como si la mano falsa formase parte de tu propio cuerpo, aun sabiendo claramente que no. Y esta experiencia «típica» varía considerablemente de una persona a otra; de hecho, muchas no sienten gran cosa. Lo mismo ocurriría también, probablemente, con las ilusiones del cuerpo entero y del intercambio de cuerpos.

Estas manipulaciones experimentales de la propiedad corporal se diferencian mucho en este sentido de las ilusiones visuales clásicas, como la del Tablero de Ajedrez de Adelson, que vimos en el capítulo 4. En el caso de las casillas de ese tablero, estamos perceptivamente tan convencidos de que son de diferentes tonos de gris que nos

sorprendemos –o incluso nos asombramos– cuando se nos revela que son en realidad del mismo tono. Esta clase de sorpresa, habitual en las ilusiones visuales, no se produce casi nunca en las ilusiones relacionadas con la propiedad corporal. En mi caso particular, la ilusión corporal más intensa que he sentido hasta ahora fue el intercambio de cuerpos que probé en Ojai, pero en ningún momento estuve siquiera cerca de creer que, de pronto, yo era otra persona o que estaba en otro lugar.

Una buena confirmación de esa debilidad subjetiva de las ilusiones de propiedad corporal la encontramos en un estudio reciente en el que participé y que examinaba el papel de la sugestionabilidad hipnótica en la ilusión de la mano de goma.[16] La lógica en la que se basaba este estudio –que dirigieron los psicólogos Peter Lush y Zoltán Dienes– era que el diseño experimental de la ilusión proporciona ya de por sí una fuerte expectativa implícita de qué se debe experimentar con ella, y que tales expectativas podrían bastar por sí solas para alterar las experiencias de propiedad corporal en algunas personas. En apoyo de esta hipótesis, descubrimos que las diferencias individuales en intensidad de la ilusión estaban correlacionadas con lo sugestionables que eran las personas según una escala estándar de hipnotizabilidad. Las personas muy hipnotizables informaban de intensas sensaciones de propiedad (durante las caricias sincrónicas), mientras que las poco hipnotizables, según la mencionada escala, apenas si eran susceptibles de sentirlas.

Por un lado, estos resultados encajan a la perfección con la hipótesis de la propiedad corporal entendida como una forma de alucinación controlada, dado que una sugestión hipnótica se puede entender como una potente expectativa generada «desde arriba», aun cuando la persona participante puede no ser consciente de tenerla.[17] Por otro lado, plantean una seria dificultad para la investigación experimental en esta área, ya que apuntan a la posibilidad de que la ilusión de la mano de goma esté causada en buena medida (o incluso por completo) por efectos de sugestión. Si los estudios sobre ilusiones de corporeización no

tienen en cuenta las diferencias individuales en sugestionabilidad –cosa que, hasta el momento, por lo general no han hecho–, es difícil que permitan afirmar nada concreto acerca de los mecanismos implicados, da igual si hablamos de manos de goma, experiencias extracorporales, ilusiones de intercambio de cuerpos o cualquier otra situación en la que (de manera implícita o explícita) se induce a las personas a esperar una experiencia corporal particular.

Es muy acusado el contraste que se aprecia entre estas ilusiones de propiedad corporal (subjetivamente leves) y las experiencias intensamente alteradas que se observan en trastornos clínicos como la somatoparafrenia, la xenomelia o el síndrome del miembro fantasma, o las intensas experiencias extracorporales asociadas a ataques o desencadenadas por estimulación cerebral directa. Estas espectaculares distorsiones se parecen mucho más a las ilusiones visuales clásicas porque concitan un convencimiento mucho mayor entre quienes tienen la experiencia inusual de turno. Y por esa razón, nos aportan pruebas mucho más contundentes de que las experiencias de corporeización y perspectiva son realmente construcciones del cerebro.

Pasemos ahora a cuestiones de identidad personal y al surgimiento el yo «narrativo» y del yo «social». Como ya vimos al hablar de la paradoja del teletransporte, esos son los niveles en los que un ente se experimenta a sí mismo como un continuo de un momento al siguiente: de un día, una semana o un mes, al día, la semana o el mes siguientes, y –hasta cierto punto– a lo largo de toda una vida. Son los niveles de yoidad en los que tiene sentido asociar el yo a un nombre, a unos recuerdos del pasado y a unos planes para el futuro. A estos niveles, cobramos conciencia de que tenemos un yo: nos volvemos verdaderamente *autoconscientes.*[18]

Estas capas superiores de la yoidad son plenamente disociables del yo corporeizado. Muchos animales no humanos, y los niños muy pequeños también, pueden experimentar una yoidad corporeizada sin tener (ni echar de menos) ningún sentido de identidad personal que la acompañe.

Y aunque los humanos adultos experimentamos normalmente todas esas formas de yoidad de un modo integrado y unificado, la disminución o la destrucción de los aspectos narrativo y social del yo pueden tener un efecto devastador en nosotros.

Clive Wearing es un musicólogo británico muy renombrado por haber editado las obras del compositor renacentista Orlande de Lassus. Fue director de coro en Londres y también remodeló el contenido musical de la emisora Radio 3 de la BBC a principios de la década de los ochenta. En marzo de 1985, en pleno apogeo de su carrera profesional, sufrió una devastadora infección cerebral, una encefalitis herpética, que causó enormes daños en el hipocampo de sus dos hemisferios cerebrales y le produjo una de las amnesias más profundas jamás documentadas.[*]

Wearing tiene unos problemas inmensos para evocar recuerdos antiguos (amnesia retrógrada), pero, sobre todo, para formar recuerdos nuevos (amnesia anterógrada). Una característica muy llamativa de su situación es que su existencia parece discurrir en una especie de presente perpetuo de entre siete y treinta segundos de duración. Hoy es ya octogenario y, probablemente, continúa experimentando la vida como una serie continua de minidespertares, como si –cada treinta segundos, más o menos– acabase de salir de un coma o de una anestesia. Tiene el yo narrativo completamente aniquilado.

El tipo de memoria que ha perdido Clive es su memoria episódica, autobiográfica: me refiero al recuerdo (episódico) de los hechos ubicados en el tiempo y el espacio, incluidos, en especial, los hechos relacionados consigo mismo (autobiográficos). Sus diarios son una lectura desgarradora.[19] Están llenos de descripciones repetidas de un «primer» despertar, una tras otra, y de frases previas (algunas escritas apenas unos instantes antes) tachadas y, en ocasiones, borradas con rabia.

8:31 a. m. Ahora estoy completamente despierto

9:06 a. m. Ahora estoy absoluta y totalmente despierto

9:34 a. m. Ahora estoy extraordinariamente despierto de verdad

Estos diarios, junto con las conversaciones con Clive transcritas por su esposa, Deborah, en su libro *Forever Today* («El hoy perpetuo»), dan fe de la agresión que padeció su sentido de identidad personal cuando sufrió aquel daño en su cerebro. Su incapacidad para hilar un relato de sí mismo a lo largo del tiempo significa que lo-que-es-ser-él ha sido un continuo empezar de cero desde hace más de treinta años, una presencia efímera sin ningún «yo» estable en torno al que organizar el flujo de percepciones del mundo y de sí mismo. Aislado en el presente por la hondura de su amnesia, su pérdida de un pasado y de un futuro es tan trastornadora para él que incluso se pregunta si está (o estaba) vivo. Deborah Wearing escribió que «Clive tenía constantemente la impresión de que acababa de salir de la inconsciencia porque no tenía prueba alguna en su propia mente de que nunca antes hubiera estado despierto […]. "No he oído nada, visto nada, tocado nada, olido nada –decía–. Es como estar muerto"».[20]

Al mismo tiempo, otros aspectos del sentido del yo de Clive se mantienen completamente intactos. No tiene problemas con las experiencias de propiedad corporal, ni con el origen de su perspectiva en primera persona, ni siquiera con la realización de acciones voluntarias. El amor que siente por su mujer no ha decaído un ápice, a pesar de que, a veces, no logra recordar haberla conocido (se casaron justo un año antes de su enfermedad). Y cuando Clive toca el piano, o canta, o dirige, la música sale libre de su interior con una fluidez que hace que parezca entero de nuevo.

Para Clive, esos momentos de amor y de música son instantes de transformación y de redención. Oliver Sacks, en un artículo muy evocador en la revista *New Yorker,* como solían ser todos los suyos, describió así la situación de Wearing: «Ya no tiene ningún relato interior; ya no lleva una vida en el mismo sentido en el que el resto de nosotros llevamos una. Pero solo hay que verlo sentado ante las teclas o

con Deborah para tener la sensación de que, en esos momentos, vuelve a ser él mismo y vuelve a estar completamente vivo».[21]

A pesar de esos momentos de gracia, la situación de Clive es a todas luces trágica. La destrucción de su yo narrativo es algo más que un simple déficit de memoria: le comporta una incapacidad para percibirse a sí mismo como un ente continuo en el tiempo, algo que viene acompañado de una erosión de su sentido fundamental de la identidad personal, ese que la mayoría de nosotros, de forma muy lógica y natural, damos por sentado. La memoria no es el principio y el fin de la yoidad, pero, como esta historia ilustra a la perfección –y como muchos sabemos por familiares y amigos perdidos en las espesuras de la demencia o de la enfermedad de Alzheimer–, es muy difícil vivir sin la persistencia y la continuidad de la autopercepción.

El poder demostrado por el amor que se profesan Clive y Deborah para restablecer el sentido de identidad de Clive nos lleva a la cuestión del «yo social».

Los seres humanos, como otros muchos animales, somos criaturas sociales. Percibir el estado mental de otra persona es una facultad crucial para las criaturas sociales en todo tipo de contextos y de sociedades. Normalmente se cree que esta aptitud –denominada a veces «teoría de la mente»– se desarrolla con bastante lentitud en los individuos humanos, pero, al final, termina desempeñando un papel clave para casi todos nosotros a lo largo de nuestras vidas.

A veces, podemos ser sumamente conscientes de esto, como, por ejemplo, cuando nos preocupa lo que nuestra pareja, un amigo o un colega pueda pensar de nosotros. Pero incluso cuando no estamos dándoles vueltas en la cabeza a nuestras interacciones sociales, nuestra capacidad para percibir las intenciones, las creencias y los deseos de otras personas siempre está interviniendo en un segundo plano, guiando nuestro comportamiento y moldeando nuestras emociones.

Existe una amplísima literatura especializada sobre la percepción social y la teoría de la mente desde ámbitos como la psicología, la

sociología y, en época más reciente, el emergente campo de la neurociencia social. Buena parte de esos artículos y libros examinan estos temas en términos de su importancia a la hora de guiar las interacciones sociales. Aquí, quisiera más bien volver el foco hacia nuestro interior para considerar cómo la experiencia de *ser yo* depende, de un modo sustancial, de cómo yo percibo a los otros que me perciben.

La percepción social –la percepción de los estados mentales de los otros individuos– no es una simple cuestión de razonamiento explícito o de «reflexión sobre» lo que otros puedan estar pensando o no.[22] Buena parte de nuestra percepción social es automática y directa. Nos formamos percepciones de las creencias, las emociones y las intenciones de otros con la misma naturalidad y facilidad con la que nos formamos percepciones de gatos, de tazas de café, de sillas e incluso de nuestros propios cuerpos. Cuando voy a servirme otra copa de vino y veo que mi amiga me acerca su copa vacía, no necesito figurarme racionalmente cuál es su intención; simplemente, percibo que le gustaría beber más vino también y que debería servirle a ella primero. Percibo estos estados mentales con la misma ausencia de esfuerzo –aunque no necesariamente con la misma precisión– con la que percibo la copa en sí.

¿Cómo ocurre esto? La respuesta, creo, radica de nuevo en la idea de que el cerebro es una máquina de predicción, y en el hecho de que la percepción es un proceso de inferencia de las causas de las señales sensoriales.

Tanto la percepción no social como la social implican que el cerebro elabore unas mejores conjeturas acerca de las causas de los *inputs* sensoriales. A veces, como bien sabemos todos, podemos malinterpretar mucho las cosas al percibir lo que se le está pasando a otra persona por la cabeza, aunque nunca confundamos una copa de vino con un coche (salvo que estemos alucinando). Uno de los motivos de la ambigüedad inherente a la percepción social es que las causas relevantes están más escondidas. Las ondas lumínicas que provocan que percibamos una copa de vino se originan de manera más o menos directa en la copa en sí,

pero las señales sensoriales relevantes de los estados mentales de otras personas deben pasar por toda una serie de etapas intermedias –de expresiones faciales, gestos y actos de habla–, cada una de las cuales da pie a una nueva oportunidad para formar una inferencia fallida.

Como en el caso de la percepción visual, las percepciones sociales dependen del contexto y la expectativa, y podemos tratar de minimizar los «errores de predicción social» tanto cambiando los datos sensoriales –mediante una variante interpersonal de la inferencia activa– como actualizando predicciones. La inferencia activa en el terreno de la percepción social equivale a comportarnos de un modo que haga cambiar el estado mental del otro individuo para que esté en sintonía con aquel que predecimos que es (o deseamos que sea).[23] Por ejemplo, sonreímos no solo para expresar nuestro propio placer, sino también para cambiar el modo en que se está sintiendo nuestro compañero o compañera de ese momento, y cuando hablamos, intentamos introducir pensamientos en la mente de la otra persona.

Estas ideas sobre la percepción social pueden vincularse con el yo social de la forma siguiente. La facultad de inferir estados mentales de otros individuos requiere –como toda inferencia perceptiva– de un modelo generativo. Los modelos generativos, como ya sabemos, son capaces de *generar* (valga la redundancia) las señales sensoriales correspondientes a una hipótesis perceptiva determinada. En el caso de la percepción social, se trata de una hipótesis acerca de los estados mentales de otra persona. Esto implica un elevado grado de reciprocidad. Mi mejor modelo de tus estados mentales incluirá un modelo sobre qué modelo te haces tú acerca de mis estados mentales. Dicho de otro modo, solo puedo entender lo que te pasa por la mente si intento comprender cómo estás percibiendo los contenidos de la mía. Es precisamente de ese modo como nos percibimos a nosotros mismos refractados a través de las mentes de otras personas. En eso consiste el yo social, y estas percepciones predictivas encajadas en lo social son una parte importante de la experiencia global de ser un yo humano.[24]

Un fascinante corolario que se deduce de esta interpretación del yo social es que es posible que la autoconciencia –los niveles superiores de la yoidad que comprenden aspectos tanto narrativos como sociales– requiera necesariamente de un contexto social para existir. Si existieras en un mundo en el que no hubiera ninguna mente más –o ninguna mente *relevante* más, para ser exactos–, tu cerebro no tendría necesidad alguna de predecir los estados mentales de otros individuos y, por consiguiente, tampoco de inferir que sus propias experiencias y acciones pertenecen a un yo. Tal vez sea literalmente verdad que «ningún hombre es una isla», como ya observara John Donne en el siglo XVII.[25]

¿Eres la misma persona que eras ayer? O tal vez sería mejor preguntarse: ¿experimentas ser tú mismo del mismo modo que lo experimentabas ayer? Probablemente (salvo que hayas sufrido algún incidente muy grave durante la noche), me dirás que sí. ¿Y si te hago esa misma pregunta con respecto a la semana pasada, el mes pasado, el año pasado, diez años atrás o cuando tenías cuatro años? ¿Y si te pregunto por cuando tengas noventa y cuatro: serás la misma persona entonces? ¿Te *parecerá* que lo eres?

Un aspecto muy curioso, aunque a menudo ignorado, de la yoidad consciente es que generalmente nos experimentemos a nosotros mismos como si fuéramos un mismo ser continuo y unificado en el tiempo. Podemos llamarlo la *estabilidad subjetiva del yo.* Hablamos no solo de una continuidad de la memoria autobiográfica, sino también de la experiencia más profunda de uno mismo o una misma como algo que persiste de un momento al siguiente, ya sea en el nivel del cuerpo biológico, ya sea en el de la identidad personal.[26]

En comparación con nuestras experiencias perceptivas del mundo exterior, nuestras experiencias relativas al yo son extraordinariamente estables. Nuestras percepciones del mundo siempre están cambiando; los objetos y escenas vienen y van en un flujo continuo de sucesos. Las experiencias relacionadas con el yo, por el contrario, parecen variar mucho menos. Aunque *sabemos* que cambiamos con el tiempo –la

mayoría disponemos de pruebas fotográficas más que suficientes de que eso es así–, igualmente nos *parece* que no cambiamos tanto. Salvo que padezcamos alguna dolencia psiquiátrica o neurológica, la experiencia de ser un yo es para nosotros muy parecida a ser un centro duradero en medio de un mundo en continua mudanza. William James, pionero decimonónico de la psicología, lo supo expresar muy bien: «A diferencia de la percepción de un objeto, que se puede percibir desde distintas perspectivas o incluso puede dejar de percibirse, con nuestro cuerpo experimentamos "la sensación del mismo cuerpo que siempre ha estado ahí"».[27]

Tal vez pienses que ahí no haya mucho en lo que ahondar; que, después de todo, los cuerpos (y otros blancos de la percepción relacionada con el yo) cambian menos que las cosas que percibimos de ahí fuera, del mundo. Yo puedo moverme de una habitación a otra, pero mi cuerpo y mis acciones y mi perspectiva en primera persona siempre me acompañan. Sobre esa base, nadie se puede sorprender de que el yo se experimente como algo que cambia menos que el mundo. Sin embargo, yo creo que sí hay mucho más que explorar sobre esta cuestión.

Como vimos en el capítulo 6, la experiencia del cambio es, en sí misma, una inferencia perceptiva. Nuestras percepciones pueden cambiar, pero eso no significa que las percibamos como cambiantes. Un buen ejemplo de esa distinción nos lo da el fenómeno de la «ceguera al cambio», por el que las cosas que varían poco a poco (en el mundo) no suscitan en nosotros una experiencia de cambio correspondiente. El mismo principio es válido también para las autopercepciones. Continuamente nos estamos convirtiendo en una persona diferente de la que éramos. Nuestras percepciones del yo cambian constantemente –ahora eres una persona ligeramente distinta de la que eras cuando empezaste a leer este capítulo–, pero eso no significa que percibamos esos cambios.

Esta ceguera subjetiva al yo cambiante tiene sus consecuencias. Para

empezar, potencia la falsa intuición de que el yo es un ente inmutable, en vez de un haz de percepciones. Pero ese no es el motivo por el que la evolución diseñó nuestras experiencias de la yoidad de ese modo. Creo que la razón de la estabilidad subjetiva del yo trasciende incluso la ceguera al cambio debida a la lentitud con la que van mudando nuestros cuerpos y nuestros cerebros. Vivimos con una forma exagerada y extrema de ceguera al autocambio, y para entender por qué, es necesario que comprendamos antes el motivo por el que nos percibimos a nosotros mismos.

Y no nos percibimos a nosotros mismos para conocernos, sino para controlarnos.

9. SER UN ANIMAL-MÁQUINA

No vemos las cosas como son,
las vemos como somos.[1]

ANAÏS NIN

El fin de la autopercepción no es descubrir lo que hay ahí fuera, en el mundo, ni aquí dentro, en el cuerpo. Su fin es el control y la regulación fisiológicos: seguir vivo. Para entender por qué lo es, y qué significa eso para *todas* nuestras experiencias conscientes, comencemos echando un vistazo a un debate muy antiguo sobre cómo están relacionadas la vida y la mente.

En la Gran Cadena del Ser –la jerarquía cristiana medieval del conjunto de la materia y de la vida–, Dios está situado en la cima.[2] Justo debajo, están los seres angelicales, luego los humanos –con las diversas subdivisiones socialmente oportunas– y, más abajo, los otros animales, las plantas y, por último, los minerales. Cada cosa está en su sitio y tiene sus propios poderes y facultades, determinados por su lugar en la Cadena.

Dentro de la Cadena, los humanos nos encontramos en un precario equilibrio entre el reino de lo espiritual, habitado por Dios y los ángeles, y el reino físico de los animales, las plantas y los minerales. Poseemos almas inmortales y la facultad tanto de razonar como de amar e imaginar, pero también somos vulnerables a las pasiones físicas –como el dolor, el hambre y el deseo sexual– porque estamos ligados a un

cuerpo.

Durante siglos, especialmente en Europa, la Gran Cadena del Ser (o *Scala Naturae)* sirvió de plantilla estable con la que los humanos podían entender su lugar en la naturaleza, además de su valor en comparación con otros congéneres: los reyes estaban más arriba en la Cadena que los campesinos, por ejemplo. Pero, en el siglo XVII, René Descartes disolvió la minuciosa gradación de la *Scala* al dividir el universo entre dos únicos modos de existencia: la *res cogitans* (o sustancia pensante) y la *res extensa* (o sustancia material).

Esta simplificada visión general de la naturaleza trajo consigo toda una serie de nuevos problemas. Había un problema de índole metafísica: ¿cómo podían interactuar esos dos mundos? Para bien o para mal (aunque mayormente para mal), esta pregunta ha definido las investigaciones sobre la conciencia desde entonces. La visión dualista cartesiana también implicaba un problema de alteraciones serias en el meticuloso orden sobre el que descansaba tanto la autoridad política como la religiosa. Si los animales tenían elementos de *res cogitans,* algún indicio de poseer una mente, ¿qué podía impedir que aspiraran a entrar en el reino espiritual como podían hacerlo los humanos? Además, todo intento de indagar intelectualmente en el alma –como aparentemente invitaba a hacer Descartes– sin duda irritaría a la poderosa Iglesia católica.[3]

Descartes siempre jugó sus cartas con cautela en lo tocante a la Iglesia, y llegó incluso a tratar de demostrar la existencia de un Dios benevolente en la tercera y la quinta de sus *Meditaciones.*[4] En cuanto a los animales no humanos, la opinión general es que él estaba convencido de que estaban totalmente desprovistos de conciencia. Pero, aunque es difícil estar seguros de ello, lo más probable es que él no pensara realmente eso.[*] En realidad, la principal tesis de Descartes acerca de los animales no humanos era que carecían de alma y de todos los atributos racionales, espirituales y conscientes asociados a la posesión de un alma. El historiador Wallace Shugg resumió así los

postulados del pensador francés a propósito de esta cuestión:

> Los cuerpos tanto de hombres como de bestias […] son meras máquinas que respiran, digieren, perciben y se mueven en virtud de la disposición de sus partes. Pero solo en el hombre dirige la razón los movimientos corporales para afrontar todas las contingencias; solo el hombre da muestras de su raciocinio usando una verdadera habla. Sin mentes que dirigen sus movimientos corporales ni reciben sensaciones, los animales deben considerarse máquinas sin pensamiento ni sentimiento que se mueven como mecanismos de relojería.[5]

Desde ese punto de vista, las propiedades carnales de los seres vivos –su naturaleza como organismos– pasan a ser irrelevantes por completo para la presencia de una mente, una conciencia o un alma (sea lo que sea esta). El mejor modo de caracterizar a los animales no humanos en ese caso es como unas *bêtes-machines,* unos «animales-máquina». En la imagen cartesiana del mundo, tan drástica es la división entre mente y vida como entre *res cogitans* y *res extensa.*

Al reforzar el carácter especial de los seres humanos, Descartes consiguió aplacar las iras de muchos de sus potenciales perseguidores. Pero dejó así una peligrosa puerta abierta. Si los animales no humanos son animales-máquina, y si los humanos somos un tipo de animal también –a fin de cuentas, parecemos ciertamente hechos de la misma clase de carne, sangre, cartílago y hueso–, entonces ¿no se explicarían las facultades de la mente y la razón en términos mecánicos y fisiológicos también?

Así lo entendió, desde luego, el filósofo francés Julien Offray de la Mettrie allá por mediados del siglo XVIII, cuando hizo extensivo el argumento del animal-máquina de Descartes a los seres humanos y defendió que las personas también eran máquinas –él habló de *l'homme machine* (el hombre máquina).[6] Con ello, negó toda condición inmaterial del alma e incluso llegó a cuestionar la existencia de Dios. La Mettrie no era de los que se andaba con chiquitas para no incomodar a la autoridad

religiosa de turno, por lo que su vida enseguida se volvió mucho más complicada que la de Descartes en su día. En 1748, se vio obligado a huir de su hogar adoptivo en los Países Bajos y entró a trabajar para el rey Federico de Prusia en su corte en Berlín, donde tres años después fallecería a raíz de un consumo excesivo de paté.

Si desde la perspectiva cartesiana la mente y la vida son independientes, para La Mettrie están estrechamente interconectadas hasta el punto de que la mente podría entenderse como una propiedad de la vida. Todavía hoy colean los argumentos y los debates sobre si la vida y la mente son sustancias continuas o discontinuas en lo que a sus mecanismos y principios subyacentes se refiere.[7]

En este debate, me alineo con las posiciones de La Mettrie, aunque lejos de hablar en términos generales sobre «la mente», yo prefiero centrar mi foco en la conciencia. Y esto nos lleva al meollo de mi teoría del «animal-máquina» sobre la conciencia y el yo. *Nuestras experiencias conscientes del mundo que nos rodea, y de nosotros mismos dentro de él, se producen con, a través de, y debido a nuestros cuerpos vivos.* Nuestra constitución animal no solo es compatible con nuestras percepciones conscientes del yo y del mundo, sino que, según postulo aquí, no podemos comprender la naturaleza y el origen de tales experiencias conscientes si no es a la luz de nuestra naturaleza como criaturas vivas.

Por debajo de las múltiples capas de expresión de la yoidad que implican recuerdos del pasado y planes para el futuro, en un estadio anterior al sentido explícito de la identidad personal, por debajo del «yo» e incluso previamente al surgimiento de una perspectiva en primera persona y de las experiencias de propiedad corporal, hay unas capas de la yoidad aún más profundas. Estas capas más cimentales están estrechamente ligadas al interior del cuerpo, más que al cuerpo como objeto en el mundo, y abarcan desde las emociones y los estados de ánimo –lo que los psicólogos llaman las experiencias «afectivas»– hasta ese sentido basal, informe y constante de simplemente «ser» un organismo viviente y corporeizado.

Iniciaremos nuestra exploración de estas profundidades con las emociones y los estados de ánimo. Estas formas de contenido consciente son fundamentales para la experiencia de ser un yo corporeizado y –como todas las percepciones– también pueden entenderse como mejores conjeturas bayesianas sobre las causas de una señal sensorial. Lo que distingue a las experiencias afectivas es que las causas relevantes se encuentran en el interior del cuerpo y no en el mundo de fuera de este.

Cuando pensamos en la percepción, tendemos a figurárnosla en términos de las diferentes formas en que sentimos el mundo exterior, sintetizadas, sobre todo, en las modalidades ya familiares de la vista, el oído, el gusto, el tacto y el olfato. Estas variedades de sensación y percepción orientadas al mundo son lo que se conoce, en conjunto, como *exterocepción.* A la percepción del cuerpo desde dentro la llamamos *interocepción,* que es el «sentido de la situación fisiológica interna del organismo».[8] * Las señales sensoriales interoceptivas se transmiten típicamente desde los órganos internos del cuerpo –las vísceras– hasta el sistema nervioso central y facilitan información sobre el estado de esos órganos, así como sobre el funcionamiento del organismo en general.[9] Las señales interoceptivas informan de cosas como los latidos del corazón, los niveles de presión arterial, aspectos diversos de bajo nivel relacionados con la química sanguínea, grados de tensión gástrica, cómo va la respiración, etcétera. Estas señales viajan por una compleja red de nervios y regiones profundas del encéfalo (en el tallo cerebral y el tálamo) hasta llegar a partes de la corteza cerebral especializadas en el procesamiento interoceptivo, como, sobre todo, la ínsula (o corteza insular).[10] * La propiedad clave de las señales interoceptivas es que reflejan, de una manera u otra, lo bien o mal que está yendo la regulación fisiológica del cuerpo, o, dicho de otro modo, lo bien que lo está haciendo el cerebro a la hora de mantener vivo el organismo en el que se encuentra.

Hace mucho que los estudiosos del tema vinculan las señales interoceptivas con la emoción y el estado de ánimo. Ya en 1884,

William James y –de forma independiente– Carl Lange postularon que las emociones no eran las «entidades psíquicas eternas y sagradas» de las que hablaban los filósofos antiguos, ni estaban tampoco «cableadas» en los circuitos cerebrales por obra y gracia de la evolución, tal como Darwin había defendido unos años antes, sino que eran percepciones de cambios en el propio estado corporal. No lloramos porque estamos tristes, sino que estamos tristes porque percibimos que nuestro estado corporal está en condiciones de llorar. Desde ese mismo punto de vista, la emoción del miedo se forma por la percepción (interoceptiva) de toda una gama de reacciones del cuerpo activadas por el organismo al reconocer un peligro en su entorno. Para James, la percepción de unos cambios físicos corporales según estos ocurren *es* la emoción: «Sentimos pena, porque lloramos; sentimos ira, porque golpeamos; sentimos miedo, porque temblamos; y no lloramos, golpeamos ni temblamos porque estemos apenados, enfadados ni asustados».[11]

La teoría de James chocó con una fuerte resistencia en aquel entonces, en parte porque subvierte la noción común, intuitiva y muy ligada a la apariencia de las cosas, de que son las emociones las que causan unas respuestas corporales, y no al revés. *Nos parece que* el sentimiento del miedo –por ejemplo, cuando nos encontramos accidentalmente con un oso gris– es el que hace que nuestro corazón se acelere, que se dispare nuestra adrenalina y que nuestros pies emprendan la huida. A estas alturas del libro, sin embargo, creo que ya hemos aprendido a ser escépticos ante la apariencia de las cosas y a no tomárnosla automáticamente como guía para entender cómo son de verdad; sería una imprudencia por nuestra parte, pues, desestimar las tesis jamesianas simplemente porque nos resulten contraintuitivas.

Más problemática para la validez de las mismas, sin embargo, es la posibilidad de que los estados corporales no sean lo bastante distintos unos de otros como para explicar toda la gama de reacciones emocionales que experimentamos los seres humanos. Aunque los detalles concretos de esta objeción son todavía objeto de debate,[12] en los

años sesenta del siglo XX surgió una potente respuesta académica y teórica con la aparición de las llamadas «teorías valorativas» de las emociones.[13] Según estas, las emociones no son más que lecturas de cambios en el estado corporal. Dependen de una valoración (o evaluación) cognitiva –es decir, a un nivel superior– del contexto en el que tienen lugar dichos cambios fisiológicos.

Las teorías valorativas resuelven el problema de la amplia diversidad emocional, porque cada emoción específica deja de corresponderse con un estado corporal concreto. Puede haber emociones muy estrechamente relacionadas –como, por ejemplo, la desgana y el hastío– que se basen en un mismo estado corporal. Las emociones resultantes serán diferentes en función de la interpretación cognitiva que se haga de esa situación corporal común. Obviamente, podría ser igualmente cierto –y sospecho que probablemente lo es– que toda emoción esté caracterizada por un marcador físico (corporal) diferenciado, y que lo que ocurra sea que los detalles menudos de sus elementos distintivos son muy difíciles de precisar.

Mi investigación experimental favorita de la teoría valorativa es un ingenioso estudio diseñado por Donald Dutton y Arthur Aron en 1974. En él, una entrevistadora abordaba a transeúntes varones que cruzaban por uno u otro de dos puentes seleccionados que unían las dos orillas del río Capilano a la altura de la localidad de North Vancouver.[14] Una de esas estructuras era un inseguro puente colgante de 135 metros de longitud, con barandillas bajas, tendido en un equilibrio bastante precario sobre un gran desnivel, por debajo del cual corrían las aguas del río formando rápidos de muy escasa profundidad. La otra era una construcción más corta y más robusta, hecha con gruesos troncos de cedro, y situada corriente arriba y a solo tres metros por encima del agua. Cuando la entrevistadora contactaba con un transeúnte que estaba cruzando uno de los puentes, lo invitaba a rellenar un cuestionario y también le ofrecía su número de teléfono, explicándole que estaría encantada de responder cualquier otra pregunta que tuviera para ella.

Lo que los investigadores se preguntaban era si los hombres que, en ese momento, estaban cruzando el puente inseguro interpretarían erróneamente como atracción sexual la excitación fisiológica causada por lo arriesgado de aquella situación, en vez de percibirla como miedo o ansiedad. Su lógica era que, de ser así, esos hombres mostrarían una mayor propensión a llamar a la entrevistadora *a posteriori* (quién sabe si incluso para pedirle una cita).

Y eso fue exactamente lo que ocurrió. La entrevistadora recibió más llamadas de hombres que estaban cruzando el puente inseguro que de los que estaban pasando por el más robusto. Dutton y Aron llamaron a lo sucedido una «mala atribución de la excitación»: los sistemas cognitivos de nivel superior de aquellos hombres habían malinterpretado el aumento de excitación fisiológica inducido por el puente inestable y lo habían atribuido a su química sexual. Como confirmación adicional de esa interpretación basada en una teoría valorativa (y presuponiendo la heterosexualidad de los encuestados), vieron también que, cuando el entrevistador era un hombre, el tipo de puente no tenía ningún efecto en el número de llamadas de seguimiento realizadas.[*]

Ese estudio, realizado hace más de cuarenta años, adolece de unos inevitables defectos metodológicos si lo comparamos con los estándares actuales, más rigurosos (aunque aún perfectibles). Y eso por no mencionar sus cuestionables aspectos éticos. Pero continúa siendo una ilustración muy elocuente de la tesis según la cual las experiencias emocionales dependen de la forma en que los procesos cognitivos de nivel superior evalúan los cambios fisiológicos.

Una limitación de las teorías valorativas es que presuponen una distinción nítida entre lo que se considera «cognitivo» y lo que no. Se asume que los sistemas perceptivos «no cognitivos», de nivel inferior, «leen» la condición fisiológica del cuerpo, al tiempo que los sistemas cognitivos (de nivel superior) «evalúan» esa condición mediante procesos más abstractos, como el razonamiento sensible al contexto. Hay

miedo, por ejemplo, cuando en primer lugar se percibe un estado corporal específico y, a continuación, se evalúa como si fuera «debido a la presencia de un oso que viene hacia mí». Sin embargo, y por desgracia para las teorías valorativas, el cerebro no establece una separación tan nítida entre el terreno de lo «cognitivo» y el de lo «no cognitivo».

Fue allá por el año 2010, en la época en que inició su andadura mi grupo de investigación en Sussex, cuando comencé a reflexionar sobre este problema. Había aprendido mucho sobre interocepción gracias a mi colega Hugo Critchley –uno de los mayores expertos mundiales en el tema– y se me ocurrió que una forma de superar las limitaciones de las teorías valorativas consistía en aplicar los principios de la predicción perceptiva y en tratar las emociones y los estados de ánimo –y las experiencias afectivas en general– como formas características de alucinación controlada.[15]

Llamé a esa idea *inferencia interoceptiva*. Del mismo modo que el cerebro carece de acceso directo a las causas de señales sensoriales exteroceptivas como las de la visión, que están en el mundo exterior, tampoco tiene un acceso directo a las causas de las señales sensoriales interoceptivas, que provienen del interior del cuerpo. *Todas* las causas de las señales sensoriales, vengan de donde vengan, le están ocultas a perpetuidad tras un velo sensorial. Así pues, en el fondo, la interocepción no deja de ser un proceso de formación de mejores conjeturas bayesianas, al igual que la percepción exteroceptiva. Del mismo modo que la «rojez» es el aspecto subjetivo de unas predicciones de base cerebral acerca de cómo reflejan la luz algunas superficies, también las emociones y los estados de ánimo son los aspectos subjetivos de otras predicciones, estas referidas a las causas de las señales interoceptivas.[16] Son unas formas de alucinación controlada inducidas desde dentro.[*]

Las predicciones interoceptivas, como las visuales, operan a múltiples escalas de tiempo y espacio, y son la base sobre la que se sustenta la

generación de «mejores conjeturas» –fluidas y sensibles al contexto y multinivel– sobre las causas de las señales interoceptivas. La inferencia interoceptiva resuelve de este modo el problema de la gran diversidad emocional sin necesidad de establecer ninguna distinción estricta entre lo no cognitivo y lo cognitivo. La inferencia interoceptiva es, por lo tanto, una explicación más directa y simple (o *parsimoniosa)* que la teoría valorativa, porque implica un solo proceso (la formación de mejores conjeturas bayesianas) en vez de dos (la percepción no cognitiva y la posterior evaluación cognitiva), y gracias a ello, también permite establecer una correspondencia más directa y fluida con el mapa de la anatomía cerebral subyacente.

La inferencia interoceptiva no es fácil de testar a nivel experimental, en parte porque las señales interoceptivas son más difíciles de medir y manejar que las exteroceptivas (las visuales, por ejemplo).[17] Un método prometedor en ese sentido es uno que explora la posibilidad de que las respuestas cerebrales a los latidos cardíacos sean tal vez marcadores distintivos de unos errores de predicción interoceptivos. La neurocientífica alemana Frederike Petzschner ha mostrado recientemente que esas respuestas, llamadas «potenciales evocados de latido cardíaco», se modulan cuando se presta atención, tal como predice la teoría de la inferencia interoceptiva. Hacen falta aún más estudios que sigan esa línea.

Otra fuente más indirecta de pruebas confirmatorias podría venir de experimentos relacionados con la propiedad corporal, como los que vimos en el capítulo anterior. En un estudio de 2013 dirigido por Keisuke Suzuki, descubrimos que las personas experimentaban mayor propiedad de una «mano de goma» de realidad virtual cuando esta destellaba en sincronía con sus propios ritmos cardíacos que cuando lo hacía de forma desacompasada, lo cual podía indicar que la propiedad corporal depende de la integración de señales tanto exteroceptivas como interoceptivas. Este método de «sincronía cardiovisual» fue utilizado también por Jane Aspell y sus colegas en un experimento de «ilusión del

cuerpo entero», en el que los participantes veían una silueta virtual de su cuerpo.[18] También estos investigadores comprobaron que las personas informaban que se identificaban más intensamente con la silueta cuando esta destellaba de forma sincronizada con los latidos de su corazón. Aunque estos estudios parecen indicar que sí existe la inferencia interoceptiva, también en este terreno se necesitan más investigaciones, en parte, porque estos experimentos no tuvieron en cuenta las diferencias individuales en sugestibilidad hipnótica, un factor que, como ya hemos aprendido, es muy importante en los experimentos sobre propiedad corporal. Sus resultados también dependen en muy buena medida de lo consciente que es cada persona de sus propios latidos cardíacos, algo que, hasta el momento, ha resultado frustrantemente difícil de medir.

Desde la perspectiva de la teoría del animal-máquina, la implicación más importante de la inferencia interoceptiva es que las predicciones interoceptivas no solo condicionan las experiencias afectivas, sino que las constituyen. Las emociones y los estados de ánimo, como todas las percepciones, proceden de dentro hacia fuera, y no de fuera hacia dentro. Tanto si hablamos del miedo, la ansiedad o la alegría, como si lo hacemos del remordimiento, toda experiencia emocional está fundada en una mejor conjetura –formada «desde arriba»– acerca del estado del cuerpo (y acerca de las causas de ese estado). Reconocer esto es el primer paso clave que debemos dar de cara a comprender hasta qué punto las experiencias de ser un yo corporeizado están ligadas a nuestra materialidad de carne y hueso.

Para dar el paso siguiente, tenemos que preguntarnos *para qué* son esas percepciones del cuerpo «desde dentro». La percepción del mundo exterior resulta a todas luces útil para guiar la acción, pero ¿por qué tiene que estar integrada desde lo más básico nuestra condición fisiológica interna en nuestras vidas conscientes? La respuesta a esa pregunta nos lleva a viajar hacia atrás en la historia una vez más, pero, en esta ocasión, solo nos remontaremos a mediados del siglo XX, y, en

concreto, a esa casi olvidada amalgama de informática, inteligencia artificial, ingeniería y biología que se bautizó con el nombre de *cibernética.*

En la década de 1950, en los albores de la era de la informática, dos disciplinas emergentes –la cibernética y la inteligencia artificial (o IA)– parecían tan prometedoras como, en muchos sentidos, inseparables. La cibernética –palabra derivada del griego *kybernetes,* es decir, «timonel» o «patrón»– fue definida por uno de sus fundadores, el matemático Norbert Wiener, como «el estudio científico del control y la comunicación en el animal y en la máquina».[19] La cibernética ponía todo su énfasis en el control y tenía su aplicación más directa en el ámbito de sistemas –como los de los misiles teledirigidos– que implicaban un bucle cerrado de realimentación entre la salida (o *output)* y la entrada (o *input)*. Una característica llamativa de este enfoque era que podía parecer que dichos sistemas tuvieran «fines» u «objetivos», como, por ejemplo, acertar en un blanco.

Esta manera de concebir las máquinas –como mecanismos potencialmente orientados a «fines»– proporcionó un nuevo nexo entre lo inerte y lo vivo. Hasta entonces, había prevalecido la idea de que solo los sistemas biológicos podían tener objetivos y, por lo tanto, solo ellos podían comportarse conforme a una finalidad interna.[*] La cibernética vino a sugerir otra cosa al poner el acento en las estrechas conexiones existentes entre máquinas y animales. En parte debido a ello, se escindió de otros enfoques de la IA, más centrados en el razonamiento desconectado (fuera de línea), incorpóreo y abstracto del que terminarían siendo ejemplos paradigmáticos los ordenadores que jugaban a ajedrez. Pero fueron estas otras aproximaciones las que, según casi todas las medidas y baremos, acabaron imponiéndose; ellas fueron las que coparon titulares de prensa y subvenciones de los organismos financiadores, mientras la cibernética se fue viendo cada vez más relegada a los márgenes del sistema. Pero aun desde ese relativo ostracismo, la cibernética produjo un buen puñado de valiosos

conocimientos cuya importancia no se ha empezado a valorar debidamente hasta fechas muy recientes.[20]

Uno de ellos salió de un artículo de William Ross Ashby y Roger Conant, publicado en 1970, en el que describen lo que llamaron el «teorema del buen regulador».[21] El concepto está muy bien sintetizado en el título de su artículo: «Todo buen regulador de un sistema debe ser un modelo de ese sistema».

Piensa en tu sistema de calefacción central o, para el caso, en tu sistema de aire acondicionado. Supongamos que está diseñado para mantener la temperatura del interior del hogar en unos inalterados 19 °C. La mayoría de sistemas de calefacción central utilizan un control de retroalimentación simple: si la temperatura es demasiado baja, se encienden, y si no, se apagan. Llamemos «Sistema A» a este tipo simple de sistema.

Imagínate ahora un sistema más avanzado, el «Sistema B». Este es capaz de *predecir* cómo responderá la temperatura dentro de la casa al hecho de que la calefacción esté encendida o apagada. Estas predicciones se basan en propiedades de la propia casa –cómo son de espaciosas sus estancias, dónde están colocados los radiadores, de qué están hechas las paredes–, pero también en cuáles son las condiciones meteorológicas en el exterior. El Sistema B ajusta entonces la salida de la caldera en el nivel correspondiente en cada momento.

Gracias a estas capacidades avanzadas, el Sistema B mantiene mejor una temperatura fija en tu casa que el Sistema A, sobre todo si tu hogar es más complejo de lo normal, o si lo es el tiempo en el exterior. El Sistema B es mejor porque tiene un *modelo* de la casa, lo que le permite predecir cómo responderá la temperatura en el interior de esta a las acciones que puede realizar. Un Sistema B de alta gama podría llegar incluso a anticiparse a desafíos térmicos inminentes –el hecho de que se avecine una jornada de frío intenso, por ejemplo– y modificar la salida de la caldera de antemano para proteger el hogar incluso frente una caída temporal del calor. Ya lo escribieron Conant y Ashby: *todo buen*

regulador de un sistema debe ser un modelo de ese sistema.[22] *

Vayamos un paso más allá con este mismo ejemplo. Imagínate que el Sistema B viene equipado con unos sensores de temperatura imperfectos, «ruidosos», que solo reflejan de forma indirecta la temperatura ambiente en el interior de la casa. Esto significa que no se puede «leer» la temperatura real directamente de los sensores, sino que debe inferirse tomando como base para ello los datos sensoriales y las expectativas previas. El Sistema B tendrá que tener entonces un modelo de: (i) cuál es la relación entre las lecturas de sus sensores y las causas ocultas de estas (la temperatura real dentro de la casa), y (ii) cómo responderán esas causas a diferentes acciones, como, por ejemplo, la de ajustar la salida de la caldera o de los radiadores.

Ahora ya podemos conectar estas ideas sobre la regulación con lo que sabemos acerca de la percepción predictiva. El Sistema B funciona infiriendo la temperatura ambiente a partir de las lecturas de sus sensores, igual que nuestro cerebro se forma mejores conjeturas posibles acerca de las causas de las señales sensoriales que recibe a fin de inferir estados del mundo (y del cuerpo) y el modo en el que cambian con el tiempo. Pero el objetivo del Sistema B no es averiguar «qué hay ahí» (en este caso, la temperatura ambiente). Su objetivo es *regular* esa causa oculta inferida y *actuar* para mantener la temperatura dentro de un intervalo confortable (y, a ser posible, sin que se salga de los límites de un único valor fijo). La percepción, entendida en ese mismo contexto, no sirve para averiguar lo que hay ahí fuera, sino para controlar y regular.

La percepción orientada al control –eso que pone en práctica el Sistema B– es, pues, una forma de *inferencia activa,* un proceso por el que los errores de predicción sensorial se minimizan realizando acciones o tomando medidas, más que actualizando predicciones. Como expliqué en el capítulo 5, la inferencia activa depende tanto de modelos generativos capaces de predecir cómo responden las causas de las señales sensoriales a las diferentes acciones, como de la modulación del equilibrio entre predicciones que proceden «desde arriba» y errores de

predicción tomados «desde abajo», para que las predicciones perceptivas puedan cumplirse realmente por sí mismas.

La inferencia activa nos indica que la percepción predictiva puede estar dirigida a inferir características del mundo (o del cuerpo) o a regularlas: puede tener por objetivo averiguar cosas o controlarlas.[23] La idea que la cibernética pone sobre la mesa es que, para algunos sistemas, el control es lo prioritario. Desde la perspectiva del «teorema del buen regulador», toda la mecánica de la percepción predictiva y la inferencia activa surge de un requisito fundamental relacionado con lo que se necesita para regular adecuadamente un sistema.

Para responder a la pregunta de para qué sirven las percepciones de las emociones y los estados de ánimo, todavía necesitamos un concepto más tomado de la cibernética: me refiero al de *variable esencial*. Introducidas también en su día por Ross Ashby, las variables esenciales con cantidades fisiológicas, como la temperatura corporal, los niveles de azúcar, los de oxígeno y otros parecidos, que se deben mantener dentro de unos límites bastante estrictos para que el organismo siga vivo.[24] Por analogía, la temperatura deseada en una habitación sería la «variable esencial» en un sistema de calefacción central.

Si unimos todas esas piezas entre sí, podemos ver por fin que las emociones y los estados de ánimo se pueden entender como unas *percepciones orientadas al control que regulan las variables esenciales del organismo*. Son *para* eso. La experiencia de miedo que siento cuando se me acerca un oso es una percepción de mi cuerpo (o, para ser más precisos, de «mi cuerpo en presencia de un oso que se le acerca») orientada al control que pone en marcha las acciones que, según las mejores predicciones, mantendrán mis variables esenciales dentro del intervalo en el que tienen que estar. Nótese que estas acciones pueden ser tanto movimientos externos del cuerpo –correr, por ejemplo– como «interoacciones» internas, como elevar el ritmo cardíaco o dilatar los vasos sanguíneos.[25]

Esta perspectiva sobre las emociones y los estados de ánimo los liga

más estrechamente aún a nuestra naturaleza de carne y hueso. Estas formas de autopercepción no tienen simplemente como finalidad registrar el estado del cuerpo desde fuera o desde dentro. Están íntima y causalmente vinculadas a lo bien que estamos y a lo bien que probablemente estaremos en el futuro: están conectadas con nuestro esfuerzo de mantenernos vivos.

Al establecer esta distinción, conseguimos también hallar la razón crucial por la que las emociones y los estados de ánimo tienen esa fenomenología característica suya. Las experiencias del miedo, los celos, la alegría y el orgullo son muy diferentes, pero son más similares entre sí de lo que cualquiera de ellas lo es con una experiencia visual o una auditiva. ¿Por qué? La naturaleza de una experiencia perceptiva depende no solo de cuál sea el blanco de la predicción correspondiente –una taza de café sobre la mesa o un corazón que se acelera, por poner dos ejemplos–, sino también del tipo de predicción que se está realizando. Las predicciones dirigidas a averiguar cosas tendrán una fenomenología muy diferente de las dirigidas a controlar cosas.

Cuando miro la taza de café que hay sobre mi mesa de trabajo, siento la fuente impresión perceptiva de un objeto tridimensional que existe independientemente de mí mismo. Esta es la fenomenología de «lo objetual», que ya presenté en el capítulo 6. Allí postulé que lo objetual surge en la experiencia visual cuando el cerebro elabora predicciones condicionales acerca de cómo cambiarían las señales visuales si se dieran diferentes acciones, como, por ejemplo, girar la taza para mostrarnos su parte de atrás. Las predicciones perceptivas, en este caso, se dirigen a averiguar qué hay ahí, y las acciones relevantes son aquellas que, como los giros o las rotaciones, predeciblemente revelarán más acerca de las causas ocultas de las señales sensoriales.

Pensemos ahora en un ejemplo más activo: atrapar una pelota de críquet. Tal vez te parezca que la mejor forma de hacerlo sería averiguar dónde va a caer la bola y correr hasta allí lo más rápido posible. Pero «averiguar qué hay ahí» no constituye una buena estrategia y, de hecho,

no es la que siguen los jugadores expertos. En realidad, deberías seguir moviéndote de tal modo que la bola siempre «parezca igual» en cierto sentido particular (concretamente, haciendo que el ángulo de elevación de tu mirada a la bola aumente, pero a un ritmo constantemente decreciente). Resulta que si sigues esta segunda estrategia –que los psicólogos llaman «cancelación de la aceleración óptica»–,[26] tienes garantizado que interceptarás la bola.[*]

Este ejemplo hace que volvamos sobre la cuestión del control. Tus acciones –y las predicciones de tu cerebro acerca de las consecuencias sensoriales de aquellas– no tienen por finalidad averiguar dónde *está* la bola. Están orientadas a controlar la *apariencia* perceptiva de la bola. Por eso, tu experiencia perceptual no revela la posición precisa de la bola en el aire, sino algo más parecido a su «atrapabilidad» a medida que corres hacia ella. La percepción en una situación así es una alucinación *controladora* en la misma medida en que es una alucinación *controlada.*

Esta idea no es nueva y tiene un rico pedigrí histórico. En la década de 1970, el psicólogo James Gibson argumentó que solemos percibir el mundo en términos de lo que él llamó «ofrecimientos» *(affordances)*.[27] Un ofrecimiento, según lo definió Gibson, es una *oportunidad para la acción* –una puerta que abrir, una bola que atrapar–, más que una representación de «cómo son las cosas» independiente de la acción. Otra teoría, también de los años setenta, aunque menos conocida que la de Gibson, hace más hincapié aún en el control. Según la «teoría del control perceptivo», de William Powers, no percibimos las cosas primero y nos comportamos de un modo determinado después, sino que, como en el ejemplo de la bola de críquet ya mencionado, nos comportamos de tal forma que terminamos percibiendo las cosas de una manera particular.[28] Pero, aunque estas teorías tempranas iban conceptualmente por el buen camino, y están en sintonía con la perspectiva de que «lo primero es la acción» que presenté en el capítulo 5 referida al cerebro, carecían de los mecanismos predictivos concretos que sí ofrece el enfoque de la percepción como alucinación controlada (o controladora). Además, se

centraban en la percepción del mundo exterior y no prestaban atención a la del interior del cuerpo.

La ansiedad no tiene un dorso, ni la tristeza tiene costados, ni la felicidad es rectangular. Las percepciones del cuerpo «desde dentro» sobre las que se erigen las experiencias afectivas no producen experiencias sobre la forma y la posición de mis diversos órganos internos (aquí, mi bazo; allí, mis riñones). No hay una fenomenología de lo objetual en este caso como si la hay cuando miramos una taza de café sobre la mesa, ni tampoco hay nada semejante a un movimiento en un marco espacial, como cuando atrapamos una bola de críquet.[29]

Las percepciones orientadas al control en las que se sustentan las emociones y los estados de ánimo tienen por finalidad predecir las consecuencias de las acciones para el mantenimiento de las variables esenciales del organismo dentro de sus niveles adecuados. Ese es el motivo por el que, en vez de experimentar las emociones como si se tratara de objetos, experimentamos lo bien o mal que va (y que es probable que vaya) nuestra situación en general. Si estoy sentado junto a la cama de mi madre en el hospital, o preparándome para huir de un oso, la forma y la calidad de mis experiencias emocionales son como son –desconsoladas, esperanzadas, histéricas, tranquilas– debido a las predicciones condicionales que mi cerebro elabora acerca de cómo pueden afectar diferentes acciones a mi condición fisiológica actual y futura.

En las capas más profundas del yo, por debajo incluso de las emociones y los estados de ánimo, se sitúa la experiencia subterránea (desde el punto de vista cognitivo), vaga y difícil de describir, del simple hecho de *ser un organismo vivo.* En ese punto, las experiencias de la yoidad afloran dentro del sentimiento desestructurado del «ser» sin más. Es ahí donde alcanzamos el núcleo central de la teoría del animal-máquina: el ya mencionado postulado de que las experiencias conscientes del mundo que nos rodea (y de nosotros mismos dentro de él) ocurren *con, a través de, y debido a* nuestros cuerpos vivos. Es en ese punto donde todas las

ideas que he estado exponiendo acerca de la percepción y el yo cobran todo su sentido. Así que vayamos paso por paso y desde el principio.

El objetivo primario de todo organismo es mantenerse con vida. Esto es verdad casi por definición: es un imperativo impuesto por la evolución. Todos los organismos vivos procuran conservar su integridad fisiológica ante los peligros y las oportunidades. Por eso existen los cerebros. La razón por la que la evolución dotó de cerebros a los organismos no fue para que pudieran escribir poesía, resolver crucigramas ni dedicarse a la neurociencia. Desde el punto de vista evolutivo, los cerebros no son «para» pensar racionalmente, comunicarse lingüísticamente o, siquiera, percibir del mundo. La razón más fundamental por la que un organismo posee un cerebro –o un sistema nervioso de cualquier tipo– es para que *lo ayude a mantenerse vivo* procurando que sus variables fisiológicas esenciales no sobrepasen los estrechos márgenes compatibles con la continuidad de su supervivencia.

Estas variables esenciales, cuya regulación efectiva resulta determinante para el estatus vital y las perspectivas futuras de un organismo, son las causas de las señales interoceptivas. Como todas las propiedades físicas, estas causas se mantienen ocultas tras un velo sensorial. Al igual que le ocurre con el mundo exterior, el cerebro no tiene un acceso directo a los estados fisiológicos del organismo, por lo que tiene que inferirlos por medio de mejores conjeturas bayesianas.

Como sucede con todas las percepciones predictivas, estas mejores conjeturas se forman a través de un proceso (de base cerebral) de minimización de los errores de predicción, que, aplicado al contexto de la interocepción, denominamos inferencia interoceptiva. Igual que con la visión y la audición –y *todas* las modalidades perceptivas–, la percepción interoceptiva es una especie de alucinación controlada.

Si la inferencia perceptiva acerca del mundo suele ir dirigida a averiguar cosas, la inferencia interoceptiva tiene como finalidad principal controlar cosas: la regulación fisiológica. La inferencia interoceptiva es un ejemplo de inferencia activa, pues, en ella, se

minimizan los errores de predicción actuando para hacer efectivas unas predicciones formadas «desde arriba», más que actualizando las predicciones en sí (aunque esto último ocurre también). Estas acciones reguladoras pueden ser externas –procurarse comida, por ejemplo– o internas (reflejos gástricos, modificaciones provisionales de la tensión arterial, etcétera).

Esta clase de control predictivo puede admitir respuestas *anticipatorias* en forma de predicciones sobre estados futuros y sobre lo que estos dependerán de unas acciones u otras. Este tipo de control anticipatorio puede ser crucial para la supervivencia. Por ejemplo, esperar a que algo como la acidez de la sangre se salga de sus límites saludables sin aplicarle una respuesta apropiada puede ser fatal. También aquí las acciones relevantes pueden ser externas, internas o de ambos tipos. Huir a la carrera de un oso antes de que te devore es un ejemplo de regulación anticipatoria externa. El incremento temporal de la presión arterial que se necesita para correr eficazmente, o incluso para levantarte de tu mesa después de haber trabajado un buen rato, es una respuesta anticipatoria interna.

En fisiología, existe un término muy útil para describir este proceso: *alostasis.*[30] La alostasis es el proceso que se sigue para alcanzar la estabilidad mediante el cambio; se diferencia, pues, de la más conocida *homeostasis,* que simplemente significa una tendencia hacia un estado de equilibrio. Desde ese punto de vista, la inferencia interoceptiva vendría a representar la regulación alostática de la condición fisiológica del organismo.

Del mismo modo que las predicciones sobre las señales sensoriales visuales son la base sobre la que se sustentan las experiencias visuales, las predicciones interoceptivas –tanto las referidas al futuro como las relativas al aquí y el ahora– son el fundamento sobre el que se apoyan las emociones y los estados de ánimo. Estas experiencias afectivas tienen su fenomenología característica a causa de la naturaleza misma (orientada al control y relacionada con el cuerpo) de las predicciones

perceptivas de las que dependen. Son alucinaciones controladoras a la par que controladas.

Pese a estar sólidamente fundamentadas en la regulación fisiológica, las emociones y los estados de ánimo se siguen experimentando mayormente como si estuvieran relacionados (al menos en parte) con cosas y situaciones que están más allá del yo, fuera del cuerpo. Cuando siento miedo, suelo tenerle miedo a alguna *cosa*. Pero los niveles más profundos de la yoidad experimentada –los de ese vago sentimiento de «ser» sin más– parecen estar exentos de cualquiera de esos referentes externos. Ese, para mí, es el verdadero estado fundamental de la yoidad consciente: una predicción perceptiva –informe, amorfa y orientada al control– sobre la condición fisiológica presente y futura del cuerpo en sí. Ahí es donde empiezas a *ser tú* y ahí es donde encontramos las conexiones más profundas entre la vida y la mente, entre nuestra naturaleza de animales-máquina y nuestro yo consciente.

El paso final y determinante en esta teoría del animal-máquina es reconocer que, desde ese punto de partida, se puede seguir todo lo demás.[31] No somos los animales-máquina de Descartes, para quien la vida no tenía relevancia alguna en lo tocante a la mente. Ocurre justamente lo contrario: *todas* nuestras percepciones y experiencias, tanto del yo como del mundo, *todas* (repito) son alucinaciones controladas y controladoras, generadas de dentro hacia fuera, y fundamentadas en la maquinaria predictiva de carne y hueso que evolucionó, que se desarrolla, y que funciona instante a instante siempre guiada por el impulso biológico básico de seguir viva.

Somos animales-máquina conscientes, de cabo a rabo.[32]

Al final del capítulo anterior, señalé que, aunque las percepciones del mundo vienen y van, las experiencias de la yoidad nos parecen estables y continuas a lo largo de múltiples escalas temporales. Ahora podemos ver que esa estabilidad subjetiva es un corolario natural de la propia teoría del animal-máquina.

Para regular de forma eficaz la condición fisiológica del cuerpo, las

creencias probabilísticas *a priori* sobre las señales interoceptivas tienen que ser de una precisión elevada para que tiendan a cumplirse por sí mismas. Este aspecto clave de la inferencia activa garantiza que las mejores conjeturas interoceptivas graviten hacia esas creencias *a priori:* es decir, hacia las regiones de viabilidad fisiológica deseadas (predichas). Así, por ejemplo, la predicción es que mi temperatura corporal se mantenga lo suficientemente constante a lo largo del tiempo y, según la inferencia activa, ese es el motivo por el que realmente se mantiene así. La experiencia del yo corpóreo como ente relativamente invariable nace directamente, por lo tanto, de esa necesidad de tener unas creencias probabilísticas *a priori* precisas (unas predicciones fuertes) sobre estados corporales estables a efectos de la regulación fisiológica. Dicho de otro modo: mientras estemos vivos, el cerebro jamás actualizará su creencia *a priori* de que espera estar vivo.*

Más aún: dado que el «cambio» en sí es un aspecto de la inferencia perceptiva, el cerebro puede atenuar las expectativas *a priori* relacionadas con la percepción de cambios en la condición del cuerpo a fin de asegurarse más aún de que las variables fisiológicas esenciales permanezcan dentro de los límites donde deben estar. Esto implica una especie de «ceguera al autocambio», un concepto que ya presentamos en el capítulo anterior. Según esto, podríamos no percibir el carácter cambiante de nuestra condición fisiológica ni siquiera en aquellos momentos en los que esté cambiando en realidad.

Si encajamos todas esas ideas entre sí, vemos que nos percibimos a nosotros mismos como si fuéramos estables a lo largo del tiempo, en parte porque tenemos la expectativa *a priori* de que nuestra condición fisiológica está limitada dentro de unos intervalos concretos (una expectativa, por cierto, que se cumple por sí misma), y en parte también porque tenemos otra expectativa *a priori* (que también se cumple por sí misma) de que esa condición no cambia. Dicho de otro modo, la eficacia de la regulación fisiológica tal vez dependa de nuestra percepción *errónea* del estado interno del cuerpo, que juzgamos más estable (y

menos cambiante) de lo que es en realidad.[33]

Curiosamente, esta proposición podría generalizarse a otros niveles superiores de la yoidad (por encima del estado fundamental de la integridad fisiológica continuada). De hecho, seremos más capaces de conservar nuestra identidad fisiológica y psicológica, en todos los niveles de la yoidad, si no nos percibimos (o no esperamos percibirnos) como seres en continuo cambio. Me percibo estable en el tiempo en todos los aspectos de ser un yo para mejorar no mi conocimiento de mí mismo, sino mi autocontrol.[34]

Como complemento a esta estabilidad subjetiva, la mayoría de nosotros también nos percibimos como «reales» la mayor parte del tiempo. Esto puede parecer obvio, pero recuerda que, en el capítulo 6, vimos que experimentar las cosas del mundo como si estas «existieran de verdad» no es prueba de que tengamos acceso perceptivo directo a una realidad objetiva, sino una propiedad fenomenológica que precisa de explicación. En aquellas páginas, postulé que, para ser de utilidad para el organismo perceptor, nuestras mejores conjeturas tienen que experimentarse como si *existieran realmente en el mundo,* y no como las construcciones cerebrales que en verdad son.

El mismo razonamiento podemos hacer con respecto al yo. De igual modo que nos parece que la silla del rincón *es realmente* roja, y que *ha pasado realmente* un minuto desde que empecé a escribir esta frase, la maquinaria predictiva de la percepción, cuando se dirige hacia nuestro interior, hace que nos parezca que *realmente hay* una esencia estable de un «yo» en el centro de todo.

Y de la misma manera que nuestras percepciones del mundo pueden carecer a veces de la fenomenología de ser reales, también el yo puede perder su realidad. La realidad experimentada (y la estabilidad subjetiva) del yo puede sufrir altibajos en períodos de enfermedad, por ejemplo, y puede verse seriamente atenuada o incluso anulada en los trastornos psiquiátricos de despersonalización. Los ejemplos más extremos de irrealidad del yo se producen en un raro delirio descrito por

primera vez en 1880 por el neurólogo francés Jules Cotard.[35] El yo corporeizado está tan ido en el delirio de Cotard que quienes lo sufren creen que no existen o que ya están muertos. Evidentemente, la *experiencia* de que el yo es irreal no significa que una supuesta «esencia del yo» haya abandonado al sujeto en casos así. Solo implica que las percepciones orientadas al control que se asocian a las capas más profundas de la regulación corporal están seriamente dañadas.[36]

Con esta postulación de la teoría del animal-máquina, no afirmo haber demostrado que la vida es necesaria para la conciencia: no digo que la carne, los huesos y las tripas –o las neuronas biológicas– posean algo especial que hace que solo las criaturas confeccionadas con esos materiales puedan tener experiencias conscientes.[37] Puede que eso sea verdad o puede que no. Nada de lo que he dicho (hasta aquí, al menos) es un argumento fuerte a favor de una opción o de la otra, ni tampoco pretendía que lo fuera. Lo que *sí* afirmo es que, para entender por qué nuestras experiencias conscientes son como son, y cómo son las experiencias que tenemos del yo y cómo se relacionan con las experiencias que tenemos del mundo, es importante valorar y comprender las raíces profundas que toda percepción tiene en la fisiología de lo vivo.

Esta reflexión sobre la base material de la conciencia nos lleva de vuelta, una vez más, al problema difícil. De hecho, la teoría del animal-máquina acelera la disolución de ese misterio aparente. Al hacer extensivo el enfoque de la alucinación controlada a las capas más profundas de la yoidad, al revelar que esa experiencia de *el-yo-como-algo-realmente-existente* es un aspecto más de la inferencia perceptiva, las intuiciones en las que está implícitamente sustentado el enfoque del problema difícil quedan más socavadas aún. En particular, la intuición (particularmente propicia al planteamiento del problema difícil) de que el yo consciente es algo que está separado en cierto modo del resto de la naturaleza –que es un observador interno inmaterial que existe realmente y mira hacia fuera, hacia un mundo material externo– pasa a

ser una simple confusión más entre lo que las cosas *parecen* y lo que *son.*

Siglos atrás, cuando Descartes y La Mettrie estaban desarrollando sus tesis sobre las relaciones entre la vida y la mente, no era el problema difícil el asunto en litigio, sino la existencia (o inexistencia) del «alma». Y lo cierto es que, aunque quizá parezca sorprendente, también se aprecian ecos del alma en esta teoría del animal-máquina. Esta alma, sin embargo, no es una esencia inmaterial, ni una síntesis espiritual de la racionalidad. El enfoque del animal-máquina relativo a la yoidad, que pone el acento en los estrechos lazos de esta con el cuerpo físico, con los ritmos persistentes de lo vivo, nos devuelve a un espacio liberado de las entelequias de una mente computacional, a un tiempo anterior a las divisiones cartesianas entre mente y materia, entre razón y no razón. Lo que podríamos llamar el «alma» desde el enfoque aquí presentado es la expresión perceptiva de una continuidad profunda entre mente y vida. Es la experiencia que tenemos cuando alcanzamos los niveles más profundos de la yoidad corporeizada –esos vagos sentimientos de «ser» sin más– como algo *realmente existente.* No parece descabellado decir que esto es un eco del alma, porque revive concepciones de esta noción eterna más antiguas aún, concepciones –como la del Ātman en el hinduismo– desde las que nuestra esencia más íntima y profunda era más aliento que pensamiento.

No somos ordenadores cognitivos; somos máquinas sintientes.

10. PEZ EN EL AGUA

En septiembre de 2007, dejé unos días Brighton y me fui a Barcelona para dar una charla en una escuela de verano sobre «ciencia, cognición y tecnología». Aunque estaba encantado de viajar a una ciudad tan hermosa, los asuntos que tuve que atender antes de mi partida me obligaron a perderme una clase magistral de tres horas impartida por el eminente neurocientífico británico Karl Friston sobre su «principio de la energía libre» y la aplicación de este a la neurociencia. (Friston ya apareció en el capítulo 5 de este libro, cuando conocimos su concepto de inferencia activa). Tenía muchas ganas de asistir a aquella lección, porque ya entonces las ideas de Friston parecían condensar –con una gran profundidad (y complejidad) matemática– algunas de mis propias reflexiones aún embrionarias acerca de la percepción predictiva y el yo.

Resignado a perderme su conferencia, pensaba que, al menos, cuando llegara allí, podría enterarme de lo que se había hablado en ella. Pero cuando, esa noche, por fin pude presentarme en la terraza del bar de una azotea donde los asistentes a las jornadas pasaban un rato de esparcimiento, me encontré con un mar de rostros que denotaban desconcierto. Karl se había subido a toda prisa a un avión de vuelta a Londres tras su clase y había dejado confusión y perplejidad detrás de sí. Tras tres horas de detalles matemáticos y neuroanatómicos, la mayoría de la gente que había escuchado su lección estaba aún más confundida al final de esta que al principio.

Parte del problema residía, al parecer, en la descomunal escala de la propuesta de Friston. Lo primero que llama la atención del principio de

la energía libre es que se trata de una idea verdaderamente descomunal. Aúna conceptos, ideas y métodos de la biología, la física, la estadística, la neurociencia, la ingeniería, el aprendizaje automático y algunos campos más. Y su aplicación no se limita ni mucho menos al cerebro. Para Friston, el principio de la energía libre explica *toda clase* de elementos de los sistemas vivos,[1] desde la autoorganización de una célula individual hasta los detalles más concretos de los cerebros y los sistemas nerviosos, la forma general y la disposición corporal de los animales, e incluso las grandes pinceladas de la evolución en sí. Es lo más próximo a una «teoría del todo» que se ha propuesto hasta el momento en biología. No es de extrañar que la gente –yo incluido– se sintiera aturdida y apabullada.

Avancemos ahora diez años en el tiempo. En 2017, mis colegas Chris Buckley, Simon McGregor y Chang-Sub Kim, y yo, publicamos por fin nuestro propio estudio sobre «el principio de la energía libre en la neurociencia» en la revista *Journal of Mathematical Psychology*.[2] Llegar a ese punto nos había llevado nueve años más de lo previsto, pero estoy contento de que perseveráramos hasta ese punto.

Al menos, *creo* estar contento, porque lo cierto es que, incluso tras ese difícil trecho recorrido, sigue habiendo algo extrañamente inescrutable en todo ello. En internet, no son pocas las publicaciones de blog que van refiriendo con regularidad los problemas de muchos expertos para entender bien las ideas de Friston. Ahí está, por ejemplo, la que Scott Alexander tituló «God help us, let's try to understand Friston on free energy» [«Que Dios nos ayude: a ver si entendemos lo que dice Friston acerca de la energía libre»].[3] Hay incluso una cuenta paródica en Twitter, @FarlKriston, en la que se publican sentencias gnómicas del tipo «Yo soy, piense lo que piense, soy. Si no fuese, ¿por qué iba yo a pensar que soy?».

Pero el principio de la energía libre vale la pena, porque su aparente inescrutabilidad se acompaña de una elegancia y una simplicidad que insinúan una unión en lo más profundo entre la vida y la mente, y con

ello, completa la teoría del animal-máquina sobre la conciencia en varios aspectos importantes.

Y como veremos, cuando se sintetiza lo bastante, el principio de la energía libre (que abreviaremos PEL) no es tan difícil de comprender en realidad.

Aparquemos ese misterioso concepto de «energía libre» por un momento y comencemos con un postulado simple acerca de lo que significa que un organismo (o de hecho, cualquier cosa) exista.

Para que una cosa exista, tiene que haber una diferencia –un límite– entre esa cosa y todo lo demás. Si no hubiera límites, no habría cosas; no habría nada.

Este límite también tiene que perdurar, porque las cosas que existen conservan su identidad en el tiempo. Si añades una gota de tinta a un vaso de agua, se dispersará enseguida y colorará el agua, pero también perderá su identidad. Sin embargo, si lo que le tiras al vaso es una gota de aceite, esta se mantendrá reconociblemente separada del agua. La gota de aceite continúa existiendo porque no se ha dispersado de manera uniforme por el agua. Al cabo de un rato, sin embargo, también perderá su identidad, igual que las rocas terminan erosionándose hasta hacerse polvo. Cosas como las gotas de aceite y las piedras existen sin lugar a dudas, porque poseen una identidad que persiste durante un tiempo (mucho tiempo, en el caso de las rocas). Pero ni las gotas de aceite ni las piedras mantienen sus límites de forma *activa:* simplemente se dispersan con la suficiente lentitud como para que notemos que existen mientras tanto.[4]

Los sistemas vivos son diferentes. Estos, a diferencia de los ejemplos anteriores, conservan activamente sus límites en el tiempo, bien moviéndose, bien (en ocasiones) simplemente creciendo. Contribuyen activamente a preservarse diferenciados de su entorno, y esa es una característica clave de lo que hace que estén vivos. El punto de partida para el PEL es que los sistemas vivos, por el simple hecho de existir, deben resistirse activamente a la dispersión de sus estados internos. Si,

en algún momento, lo único que queda de ti es un charco de puré indistinguible en el suelo, es que ya no estás vivo.*

Esta concepción de la vida nos retrae aquí el concepto de *entropía.* En el capítulo 2, presenté la entropía diciendo que se trataba de un indicador de desorden, diversidad o incertidumbre. Cuanto más desordenados están los estados de un sistema (como esa gota de tinta que se dispersa sin orden ni concierto por toda el agua), más alta es la entropía. En tu caso, en el mío o incluso en el de una bacteria, los estados internos están menos desordenados cuando estamos vivos que cuando nos descomponemos hasta hacernos puré. Estar vivo significa encontrarse en una condición de *entropía baja.*

Y he ahí el problema. En física, la segunda ley de la termodinámica nos dice que la entropía de cualquier sistema físico aislado se incrementa con el tiempo. Todos los sistemas de ese tipo tienden al desorden, es decir, a la dispersión de sus estados constituyentes a medida que pasa el tiempo. La segunda ley nos dice que los casos de materia organizada –como son los sistemas vivos– son intrínsecamente improbables e inestables, y que (a la larga) todos estamos condenados a desorganizarnos. Pero, por la razón que sea, a diferencia de las rocas o las gotas de tinta, los sistemas vivos eluden temporalmente la segunda ley y persisten en mantener una condición precaria de improbabilidad. Existen a partir de un equilibrio con su entorno (que es justo lo que, de entrada, significa «existir»).

Según el PEL, para que un sistema vivo resista al tirón gravitacional de la segunda ley, *debe ocupar estados en los que espera estar.* Como buen bayesiano que soy, utilizo el verbo «esperar» en su sentido estadístico, no en el psicológico. Se trata de una idea muy simple, casi trivial. Un pez en el agua se encuentra en un estado en el que espera (estadísticamente) estar, porque la mayoría de peces se encuentran de hecho en el agua durante la mayor parte del tiempo. Lo estadísticamente inesperado es que un pez esté fuera del medio acuático, a menos que esté empezando su proceso de transformación en puré. Asimismo, el

hecho de que mi temperatura corporal se mantenga en torno a los 37 °C también es un valor esperado (en sentido estadístico), compatible con la continuación de mi supervivencia y con el hecho de que no comience a transformarme en papilla.

Para cualquier sistema vivo, la condición de «estar vivo» significa la búsqueda proactiva de un conjunto particular de estados a los que se regresa reiteradamente a lo largo del tiempo, ya sean temperaturas corporales, ritmos cardíacos (las «variables esenciales» fisiológicas que ya vimos en el capítulo anterior) o la organización de complejos proteínicos y flujos de energía dentro de una bacteria unicelular. Estos son los estados de baja entropía estadísticamente esperados que garantizan que el sistema se mantenga con vida: son, en definitiva, el tipo de estados esperados dada la criatura en cuestión.*

Es importante que tengamos en cuenta que los sistemas vivos no son sistemas cerrados y aislados. Los sistemas vivos están en continua interacción abierta con su entorno, del que cosechan recursos, nutrientes e información. Aprovechando esa apertura es como los sistemas vivos son capaces de dedicarse a la actividad (muy consumidora de energía) de buscar estados estadísticamente esperados, minimizar la entropía y eludir la segunda ley.

Desde la perspectiva de un organismo, la entropía que importa es la de sus estados *sensoriales,* que son aquellos que lo ponen en contacto con su entorno. Imagina un sistema vivo muy simple: una bacteria individual, por ejemplo. Esta bacteria necesita un nutriente concreto para sobrevivir y puede percibir la concentración del mismo en su entorno inmediato. *Esperando* percibir elevadas concentraciones del nutriente y buscando activamente (mediante sus movimientos) encontrar las señales sensoriales esperadas, este organismo simple se mantendrá dentro del conjunto de estados que lo definen como ente vivo. Dicho de otro modo, percibir altas concentraciones de nutriente es un estado estadísticamente esperado para la bacteria, un estado al que trata proactivamente de regresar.

Según el PEL, este comportamiento es de aplicación general. En última instancia, *todos* los organismos –no solo las bacterias– se mantienen con vida minimizando su entropía sensorial a lo largo del tiempo y contribuyendo de ese modo a asegurarse de que permanecen dentro de los estados estadísticamente esperados que son compatibles con la supervivencia.

Es en este punto donde llegamos al núcleo central del PEL, aquel en el que se aborda la cuestión de cómo consiguen los sistemas vivos, en la práctica, minimizar su entropía sensorial. Por lo normal, para minimizar una cantidad, un sistema tiene que ser capaz de medirla. El problema en este caso es que la entropía sensorial no es algo que se pueda detectar ni medir de forma directa. Un sistema no puede «saber» si sus propias sensaciones son sorprendentes basándose simplemente en las sensaciones en sí. (Veámoslo con una analogía. ¿Es sorprendente el número 6? Es imposible decirlo sin conocer el contexto). De ahí que la entropía sensorial sea muy diferente de cosas como los niveles lumínicos, o las concentraciones de nutrientes cercanos, que un organismo sí puede detectar directamente por medio de sus sentidos y utilizar para guiar su conducta.

Y aquí es donde, por fin, la *energía libre* entra en escena. Que nadie se inquiete con el nombre, cuyos orígenes se remontan a las teorías decimonónicas de la termodinámica.[*] A los efectos que aquí nos ocupan, podemos concebir la energía libre como una magnitud que se aproxima a la entropía sensorial.[5] Lo crucial, en este caso, es que es también una cantidad que un organismo puede medir (y, por lo tanto, que puede minimizar).

Siguiendo la lógica del PEL, podemos decir entonces que los organismos se mantienen a sí mismos en los estados de entropía baja que les garantizan la continuación de su existencia minimizando activamente esa cantidad medible llamada energía libre. Pero ¿qué es la energía libre desde la perspectiva del organismo? Al final, y tras ciertos malabarismos matemáticos, resulta que la energía es básicamente lo

mismo que el *error de predicción* sensorial.[6] Cuando un organismo está minimizando dicho error, como se prevé que haga desde teorías como la del procesamiento predictivo o la de la inferencia activa, también está minimizando su cantidad de energía libre (por mucha mayor profundidad teórica que este último concepto tenga).

Una de las implicaciones de esta conexión es que el PEL permite retomar la idea, presentada en el capítulo anterior, de que los sistemas vivos tienen (o son) modelos de su entorno. (Concretamente, modelos de las causas de sus señales sensoriales). Y lo permite porque, en el procesamiento predictivo, como vimos en el capítulo 5, se necesitan modelos que suministren las predicciones que, a su vez, definen los errores de predicción. Según el PEL, poseer o ser un modelo es lo que hace que un sistema pueda juzgar si sus sensaciones son (estadísticamente) sorprendentes o no. (Si crees que el número 6 que estás viendo proviene del lanzamiento de un dado, puedes valorar exactamente lo sorprendente que es).

Estas conexiones profundas entre el PEL y el procesamiento predictivo tienen bastante sentido (y muy atractivo, por cierto). Podemos hacernos una idea bastante intuitiva, por ejemplo, de cómo, al minimizar el error de predicción por medio de la inferencia activa, los sistemas vivos llegarán a estar naturalmente en aquellos estados en los que esperan (o predicen) estar. Así entendidas, las ideas de la percepción predictiva y la alucinación controlada (o controladora) se siguen bastante directamente del ambicioso intento de Friston de explicar la totalidad de la biología.

Cuando reunimos todos estos puntos, la imagen que surge es la de un sistema vivo que elabora activamente modelos sobre su mundo y su cuerpo para poder regresar, una y otra vez, a ese conjunto de estados que lo definen como tal sistema vivo: desde el latido de mi corazón segundo a segundo hasta lamentarme cada doce meses de que sea el día de mi cumpleaños. Parafraseando a Friston, la perspectiva que brinda el PEL es la de unos organismos que recogen información sensorial y elaboran modelos con ella para maximizar las pruebas sensoriales de su

propia existencia.[7] O, como a mí me gusta decir: «Me predigo, *ergo sum*».

Vale la pena señalar que minimizar la energía libre –el error de predicción sensorial– *no* significa que un sistema vivo pueda seguir siéndolo si se retira a una habitación oscura y silenciosa y se queda allí, mirando fijamente la pared sin hacer otra cosa. Aunque pueda pensarse que esta sería una estrategia ideal, pues, en un caso así, los *inputs* sensoriales del entorno externo serán sumamente predecibles, de ideal tiene muy poco. Con el tiempo, los *inputs* sensoriales que señalan otras cosas, como los niveles de azúcar en sangre y otras por el estilo, comenzarían a desviarse de sus valores esperados: si te quedas en esa habitación oscura demasiado tiempo, acabarás pasando hambre. La entropía sensorial empezaría a crecer y comenzaría a rondarte el fantasma de una inminente inexistencia. Los sistemas complejos como son los organismos vivos necesitan dejar que cambien algunas cosas para que otras se mantengan igual. Tenemos que movernos para levantarnos de la cama y preparar el desayuno, y nuestra presión arterial tiene que aumentar mientras lo hacemos para que no nos desmayemos. Esto se corresponde con esa forma anticipatoria de control predictivo –la alostasis– que mencioné en el capítulo previo. Minimizar el error de predicción sensorial a largo plazo significa salir de la habitación oscura o, cuando menos, encender la luz.[8]

Otra de las preocupaciones habituales con el PEL es que no sea falsable, en el sentido de que no se pueda contrastar ni refutar con datos experimentales. Esto es cierto, pero no es una característica privativa del PEL ni particularmente problemática. En el fondo, el PEL es más una filosofía matemática que una teoría concreta que se pueda evaluar mediante contrastación de hipótesis. Tal como lo plantea mi colega Jakob Hohwy, el PEL trata de responder a la pregunta de «¿cuáles son las condiciones para la posibilidad de la existencia?» siguiendo el mismo razonamiento a partir de primeros principios con el que Immanuel Kant se formuló la pregunta de «¿cuáles son las condiciones para la posibilidad de la percepción?». El papel del PEL puede concebirse como

uno de motivación y facilitación de la interpretación de otras teorías más específicas, teorías susceptibles (estas sí) de refutación experimental. La teoría del procesamiento predictivo, por ejemplo, puede falsarse si se comprueba que el cerebro no usa errores de predicción sensorial en el proceso de percibir. Al final, el PEL se evaluará por lo útil que resulte, y no por si es empíricamente verdadero o falso en sí mismo.[9*]

Resumamos, pues, los principales pasos del PEL. Para que los organismos se mantengan vivos, necesitan comportarse de un modo que les permita permanecer en los estados (de baja entropía) en los que «esperan» estar. Un pez nadando por encima de un arrecife de coral en busca de alimento está buscando de forma proactiva estados sensoriales compatibles con la continuación de su supervivencia. En general, los sistemas vivos hacen esto minimizando una aproximación medible de la entropía de esos estados, que es la energía libre. Minimizar la energía libre exige que el organismo tenga (o sea) un modelo de su entorno, en el que está incluido su cuerpo. Los organismos minimizadores de energía libre usan esos modelos para reducir la diferencia entre las señales sensoriales predichas y las reales, para lo cual actualizan predicciones y realizan acciones. De hecho, si se aceptan una serie de verosímiles supuestos, la energía libre resulta ser exactamente lo mismo que el error de predicción. En general, eso significa que el conjunto del procesamiento predictivo, las alucinaciones controladas, la inferencia activa y la percepción orientada al control (y de la teoría del animal-máquina, también) se puede concebir –a través de la óptica del PEL– como una derivación natural de cierta condición restrictiva fundamental de lo que significa estar vivo, de lo que significa existir.

Por si acaso este rápido repaso del PEL resulta un tanto desorientador, quisiera decir para tranquilizar al lector que no hace falta comprender ni aceptar el PEL para seguir el argumento teórico de las alucinaciones controladas y los animales-máquina que he venido exponiendo en los capítulos anteriores.[*] La teoría según la cual experimentamos el mundo

(y el yo) a través de unos mecanismos de percepción predictiva que están fundamentados en un «impulso por seguir vivo» se sostiene por sí sola. Sin embargo, merece la pena introducir este excurso por el PEL porque mejora la teoría del animal-máquina en, por lo menos, tres aspectos significativos.[11]

En primer lugar, el PEL ancla la teoría del animal-máquina en la rocosa base de la física y, en particular, de una física relevante para lo que significa estar vivo. El «impulso de seguir vivo» aflora de nuevo en el PEL, pero en forma de un imperativo más fundamental aún: el de mantenerse dentro de unos estados estadísticamente esperados; el de resistir al insistente tirón gravitacional de la segunda ley de la termodinámica. Cuando una teoría puede generalizarse y fundamentarse de este modo, se vuelve más convincente, más integradora y más potente.[11]

En segundo lugar, el PEL concreta y fortalece la teoría del animal-máquina porque nos la cuenta en sentido inverso. En los capítulos previos, comenzamos planteándonos el reto de lo que supone inferir cómo es el mundo exterior desde el interior del cráneo, una verdadera cámara acorazada ósea, y fuimos siguiendo el hilo de ideas desde fuera hacia dentro, hasta llegar al interior del cuerpo (empezando por las experiencias de la yoidad entendidas como mejores conjeturas perceptivas, y terminando por identificar las más profundas de esas experiencias como un ejercicio de percepción del cuerpo en sí orientada al control). Con el PEL, el proceso es a la inversa. Comenzamos con la simple afirmación de que «las cosas existen» y procedemos desde ahí hacia fuera: hacia el cuerpo y hacia el mundo. Llegar al mismo lugar desde dos puntos de partida muy diferentes refuerza la intuición de que el argumento que subyace es coherente, y deja claros ciertos paralelismos entre conceptos (como, por ejemplo, entre la energía libre y el error de predicción) que, de otro modo, serían de más difícil comprensión.

La tercera ventaja del PEL radica en la variada caja de herramientas

matemáticas que pone sobre la mesa. Este kit ofrece múltiples oportunidades nuevas de desarrollar aún más las ideas que he presentado en capítulos anteriores. He aquí un ejemplo. Cuando desplegamos las matemáticas del PEL en mayor detalle, descubrimos que lo que de verdad necesito para seguir vivo es minimizar la energía libre *en el futuro,* y no solo aquí y ahora. Y resulta que minimizar este error de predicción a largo plazo significa que tengo que buscar nuevas sensaciones *ahora* que reduzcan mi incertidumbre acerca de lo que ocurriría *a continuación,* si hiciera esto o lo otro. Me convierto así en un agente curioso, buscador de sensaciones, y no en alguien que se conforma con confinarse en una habitación oscura. Las matemáticas del PEL ayudan a cuantificar ese tenue equilibrio entre exploración y aprovechamiento, y esto, a su vez, tiene sus implicaciones de cara a lo que percibimos, ya que lo que percibimos se construye en todo momento y lugar con las predicciones que elabora el cerebro. Ideas como esta nos permiten diseñar mejores experimentos, tender puentes explicativos más sólidos que soporten el peso de esos experimentos y, poco a poco, puente a puente, nos va acercando a una explicación satisfactoria de cómo los mecanismos dan origen a las mentes.[12]

Pero, por otra parte, y pese a que el PEL se anuncia como una «teoría del todo», no es una teoría de la conciencia. El PEL guarda la misma relación con la conciencia que las teorías bayesianas predictivas del cerebro: son teorías *para* la ciencia de la conciencia, en el sentido del problema real, pero no *de* la conciencia, en el sentido del problema difícil.[13] El PEL aporta nuevas intuiciones y herramientas al desafío de explicar la fenomenología en términos de mecanismos. Y a cambio, las nociones de las alucinaciones controladas y de los animales-máquina dotan las austeras matemáticas del PEL de cierta relevancia recién descubierta para la conciencia. Además, ¿de qué sirve una teoría del todo si no tiene algo que decir al respecto?

Muchos años después de mi primer (y desconcertante) encontronazo con el PEL, pasé unos días con Karl Friston –y una veintena de otros

neurocientíficos, filósofos y físicos– en un pequeño simposio en la isla griega de Egina, a una hora en ferri desde Atenas. Fue otro septiembre, este de 2018, no mucho después de la singladura de mi madre por los mares del delirio. Como en el viaje a Barcelona de más de una década antes, fui allí con ganas de disfrutar de un poco de ciencia y también de un poco del sol de finales del verano. El plan previsto era hablar y debatir sobre el PEL, aunque centrándonos en la relación de este con la teoría de la información integrada (TII) referida a la conciencia, la no menos ambiciosa construcción teórica que examinamos en el capítulo 3. Sin embargo, en lugar de cálido sol y cielos despejados, nos recibió una impresionante tormenta, un inusitado medicán que arrastró mesas y sillas hacia el mar y agitó el Mediterráneo, tan plácido en condiciones normales, hasta convertirlo en una furia de aguas bravas.

Sentados en el anexo donde celebrábamos el encuentro, mientras las ráfagas de viento y lluvia hacían batir las puertas y movían descontroladas las ramas, que fustigaban las ventanas, me di cuenta de lo extraordinario que resultaba estar trabajando sobre la conciencia en una época en la que contábamos con dos teorías muy ambiciosas y muy detalladas en el plano matemático que no parecían interpelarse mutuamente en lo más mínimo. De entrada, esa ausencia de interacción podría considerarse desalentadora, pero a mí más bien me parecía una situación fascinante.

La tormenta continuó azotándonos todo el día. Se plantearon algunas ideas, pero tuve la sensación de que andábamos básicamente buscándolas sin encontrarlas, sumidos en una especie de semioscuridad. El PEL y la TII son «grandes» teorías, pero lo son en sentidos distintos. El PEL parte de un postulado simple («las cosas existen») para derivar de él la totalidad de la neurociencia y la biología, pero no la conciencia. La TII parte de otro enunciado simple («la conciencia existe») y lanza a partir de él un asalto directo al problema difícil. No es de extrañar que tiendan a no escucharse cuando hablan.

Dos años después, cuando le estoy dando los últimos retoques a este

libro, sigo viendo que las dos teorías viven en mundos diferentes. Pero ahora, al menos, hay algunas tentativas en marcha dirigidas a comparar sus respectivas predicciones experimentales.[14] Los debates de planificación de dichos experimentos –de los que he tenido la suerte de ser partícipe– han sido, a ratos, tan ilustrativos como frustrantes debido, sobre todo, a lo radicalmente distintos que son los puntos de partida y los objetivos explicativos que cada teoría pone sobre la mesa. Todavía está por ver qué resultará de esos experimentos. Intuyo que aprenderemos muchas cosas útiles, pero que ni el PEL ni la TII quedarán explícitamente descartadas como teorías de (ni para) la conciencia.

Mis propias ideas sobre alucinaciones controladas y animales-máquina trazan una especie de camino intermedio. Comparten con el PEL una profunda fundamentación teórica en la naturaleza del yo y se valen de la potente maquinaria matemática y conceptual del cerebro predictivo. Por su parte, tienen en común con la TII su clara focalización en las propiedades subjetivas, fenomenológicas, de la conciencia, aunque con el problema real (no el difícil) en su punto de mira. Más que para confrontar el PEL con la TII, lo que espero es que la teoría del animal-máquina sobre la conciencia y el yo sirva para hallar el modo de reunirlas, de entretejer ideas de ambas para confeccionar una imagen satisfactoria de por qué somos lo que somos.

El encuentro en Egina terminó sin especial algarabía, como suele ocurrir en estos casos. Cuando subimos al ferri que nos devolvió a Atenas, la tormenta ya había amainado y el mar estaba en calma. Me había costado dar el paso de hacer ese viaje, pero, al final, lo di, y allí de pie al sol, en la cubierta del transbordador, sentí que había hecho bien. Comencé a pensar en cómo había tomado la decisión, en por qué siempre cuesta tanto, y enseguida me puse a cavilar también sobre cómo decidimos, sea lo que sea que decidamos, y sobre lo que significa que sintamos que controlamos nuestras elecciones o nuestros comportamientos en general.

Y es que, en cuanto empiezas a pensar en el libre albedrío, ya no hay

manera de parar.

11. GRADOS DE LIBERTAD

> [Ella] dobló el dedo y lo enderezó. El misterio estaba en el instante antes de que se moviese, en la línea divisoria entre el no moverse y el moverse, cuando su intención surtía efecto. Si pudiera estar en la cima, pensó, quizás descubriese el secreto de sí misma, aquella parte de sí que mandaba en realidad. Acercó el índice a la cara y lo miró fijamente, instándole a moverse. Permaneció inmóvil porque ella estaba simulando […]. Y cuando finalmente dobló el dedo, pareció que la acción empezaba en el propio dedo, no en alguna parte de [su] mente.[1]
>
> IAN MCEWAN, *Expiación*

¿Cuál es el aspecto de *ser tú* al que te aferras con mayor firmeza? Para muchas personas es esa sensación de tener control sobre sus actos o de ser dueñas de sus pensamientos. Hablamos de la cautivadora, aunque ciertamente compleja, idea de que actuamos conforme a nuestro libre albedrío.

Ian McEwan detecta esa complejidad incluso en el simple acto de doblar un dedo. Briony Tallis siente a sus trece años que sus intenciones conscientes (como, por ejemplo, la de doblar un dedo) causan unas acciones físicas (como que el dedo se doble de verdad). La línea de causación aparente va directa de la intención consciente a la acción física. Y ella siente que, en ese proceso, radica la propia esencia de la yoidad, de lo que significar ser ella. Pero cuando Briony ahonda un poco más en esas sensaciones, las cosas dejan de ser tan simples. ¿Dónde se

originó el movimiento? ¿En la mente o en el dedo? ¿Fue la intención –de su «yo»– la que causó la acción, o fue la experiencia de la intención una consecuencia de percibir que el dedo comenzaba a moverse?

Briony Tallis no es ni mucho menos la única que ha reflexionado sobre esas cuestiones. Pocos temas han sido objeto de tan constante y encendido debate en la filosofía y en la neurociencia como el libre albedrío. Qué es, si existe o no, cómo se produce, si importa o no: nunca ha sido fácil encontrar un mínimo de consenso a propósito de ninguna de esas cuestiones. Ni siquiera está clara la *experiencia* en sí del libre albedrío: es decir, si es una experiencia unitaria o si se trata más bien de una categoría de experiencias interrelacionadas, o si difiere según la persona, etcétera.[2] Pero entre tanta confusión hay una intuición que sí se ha mantenido estable. Cuando ejercemos el libre albedrío, tenemos –por citar al filósofo Galen Strawson– una sensación de «radical, absoluta y autorresponsabilizadora *depende-de-mí-dad* en lo tocante a nuestro elegir y nuestro actuar».[3] Es la sensación de que el yo está desempeñando un papel causal en la acción como no lo tendría en el caso de una reacción meramente refleja, como cuando retiras la mano ante la picadura de una ortiga. De ahí que la experiencia del libre albedrío acompañe de manera natural a las acciones voluntarias, tanto si se trata de doblar un dedo como de prepararse una taza de té o emprender una nueva carrera profesional.

Cuando experimento que realizo una acción por mi «libre voluntad», en cierto sentido estoy experimentando que es *mi yo* la causa de dicha acción. Las experiencias de volición hacen (posiblemente más que ningún otro tipo de experiencia) que sintamos que hay un «yo» inmaterial consciente moviendo hilos en el mundo material. Así parecen las cosas.

Pero las experiencias de volición no revelan la existencia de un yo inmaterial con poder causal sobre los sucesos físicos. Yo creo, más bien, que constituyen formas características de percepción relacionada con el yo. Para ser más precisos, son percepciones relativas al yo asociadas a

acciones voluntarias. Como todas las percepciones –tanto las relativas al yo como las relativas al mundo–, las experiencias de volición se construyen con arreglo a los principios de elaboración de mejores conjeturas bayesianas, y desempeñan papeles importantes (o probablemente esenciales) de cara a guiar lo que hacemos.

Aclaremos, primero, lo que el libre albedrío *no* es. El libre albedrío no es una intervención en el flujo de los sucesos físicos en el universo (o, para ser más concretos, en el cerebro) que hace que ocurran cosas que, de otro modo, no ocurrirían. Este libre albedrío «espectral» evoca el dualismo cartesiano, exige toda liberación de las leyes de la causa y el efecto, y a cambio, no aporta nada que tenga un valor explicativo.

Descartar ese libre albedrío espectral implica que también podamos desechar ese persistente –aunque equivocado– interés por saber si el *determinismo* es real o no. En física y en filosofía, el determinismo significa suponer que todo lo que acontece en el universo está completamente determinado por causas físicas preexistentes. La alternativa al determinismo es que el universo lleve incorporado el azar desde sus propios fundamentos, ya sea a través de fluctuaciones en una sopa cuántica, ya sea a través de algún otro principio de la física aún desconocido. La importancia del determinismo para el libre albedrío ha sido el tema de interminables debates. Mi antiguo jefe, Gerald Edelman, lo supo resumir muy bien en una única y provocadora frase: el libre albedrío –se piense lo que se piense acerca de él– es algo que estamos determinados a tener.

En cuanto sacamos el libre albedrío espectral de la ecuación, el debate sobre el determinismo pierde toda su importancia. Ya no hay necesidad alguna de dejar cierto margen de maniobra no determinista para que esa libre voluntad nuestra pueda intervenir. Desde la perspectiva del libre albedrío como experiencia perceptiva (que yo propugno aquí), simplemente deja de ser necesario que se produzca alteración alguna en el flujo causal de los sucesos físicos. El universo determinista puede seguir tranquilamente con su lento devenir.[4] Y si el determinismo es

falso, no importa lo más mínimo, porque el ejercicio de la libre voluntad no significa comportarse de forma aleatoria. Ni *sentimos* las acciones voluntarias como algo azaroso ni tampoco *son* aleatorias.[5]

A principios de la década de 1980, en la Universidad de California en San Francisco, el neurocientífico Benjamin Libet realizó una serie de experimentos sobre la base cerebral de la acción voluntaria que no han dejado de suscitar controversia desde entonces. Libet aprovechó un conocido fenómeno llamado «potencial de disposición» (o de preparación), que es una pequeña señal de EEG con forma de pendiente que se origina en algún punto situado por encima de la corteza motora y precede de manera sistemática a las acciones voluntarias.[6] Libet quería saber si esa señal cerebral se podía detectar con anterioridad, no ya a una acción voluntaria, sino también al momento en que la persona fuese siquiera *consciente de la intención* de realizar aquella acción.

Su diseño experimental, mostrado en la imagen de la página 256, fue simple y directo. Libet pidió a los participantes que doblaran la muñeca de su brazo dominante en el momento que ellos mismos eligieran a fin de que realizaran una acción voluntaria espontánea (como la de Briony en la novela de McEwan). Cada vez que lo hacían, él medía el momento exacto del movimiento a la vez que usaba el EEG para registrar la actividad cerebral tanto anterior como posterior al inicio de este. También pidió a sus voluntarios (y esta era la parte crucial de su ensayo) que estimaran cuándo experimentaban el «impulso» de hacer cada movimiento: el momento preciso de la intención consciente, la cresta de la ola justo antes de que esta rompa.[7] Para ello señalaban la posición angular de un punto rotatorio sobre la pantalla de un osciloscopio en el instante en que experimentaban la intención de moverse, y luego informaban esa posición.

Los datos fueron claros. Tras promediar muchos ensayos, se vio que el potencial de disposición era detectable unos cientos de milisegundos *antes* de la intención consciente de mover la muñeca. Dicho de otro modo, para cuando una persona es consciente de su intención, el

potencial de disposición ya ha empezado a intensificarse.[8]

Una interpretación habitual del experimento de Libet es que «refuta el libre albedrío». En realidad, para el que es sin duda muy mala noticia (en un mundo lleno de estas) es para el libre albedrío espectral, porque parece excluir la posibilidad de que la experiencia de la volición causara la acción voluntaria. Al propio Libet le preocupaba lo bastante esa implicación como para emprender lo que, visto desde nuestra perspectiva actual, pareció un intento desesperado de rescate: concretamente, lanzó la idea de que, entre el momento del impulso y la acción resultante, quedaba tiempo suficiente para que interviniera el libre albedrío espectral e *impidiera* que la acción tuviera lugar. Si no hay un libre albedrío auténtico (entiéndase espectral), pensó Libet, tal vez sí siga habiendo una «libre voluntad de no hacer».[9] Es un buen intento, pero, obviamente, no cuela. La *inhibición* consciente no tiene menos de milagrosa que la *intención* consciente.

Hace décadas que se debate sobre qué *implicaciones* concretas tuvieron las observaciones de Libet para el libre albedrío. Parece raro que el potencial de disposición se pueda identificar con tanta antelación respecto a la acción voluntaria. En cronología cerebral, medio segundo es muchísimo tiempo. De hecho, hubo que esperar a 2012 para que una nueva idea y un inteligente experimento revolucionaran adecuadamente la situación: fue cuando el neurocientífico Aaron Schurger se dio cuenta de que los potenciales de disposición podrían no ser marcadores del hecho de que el cerebro estuviera iniciando una acción, sino más bien productos del modo en que se miden.

Los potenciales de disposición se miden típicamente mirando hacia atrás en el tiempo, en el EEG, comenzando a partir de esos momentos en los que una acción voluntaria ocurrió realmente. Schurger se percató de que, obrando de ese modo, los investigadores ignoran sistemáticamente todos los demás momentos en que *no* se producen acciones voluntarias. ¿Cómo se vería la EEG en esos otros instantes? ¿Es posible que haya una actividad similar a los potenciales de disposición que esté operando todo

el tiempo, pero que no veamos porque no es lo que andamos buscando? [10]

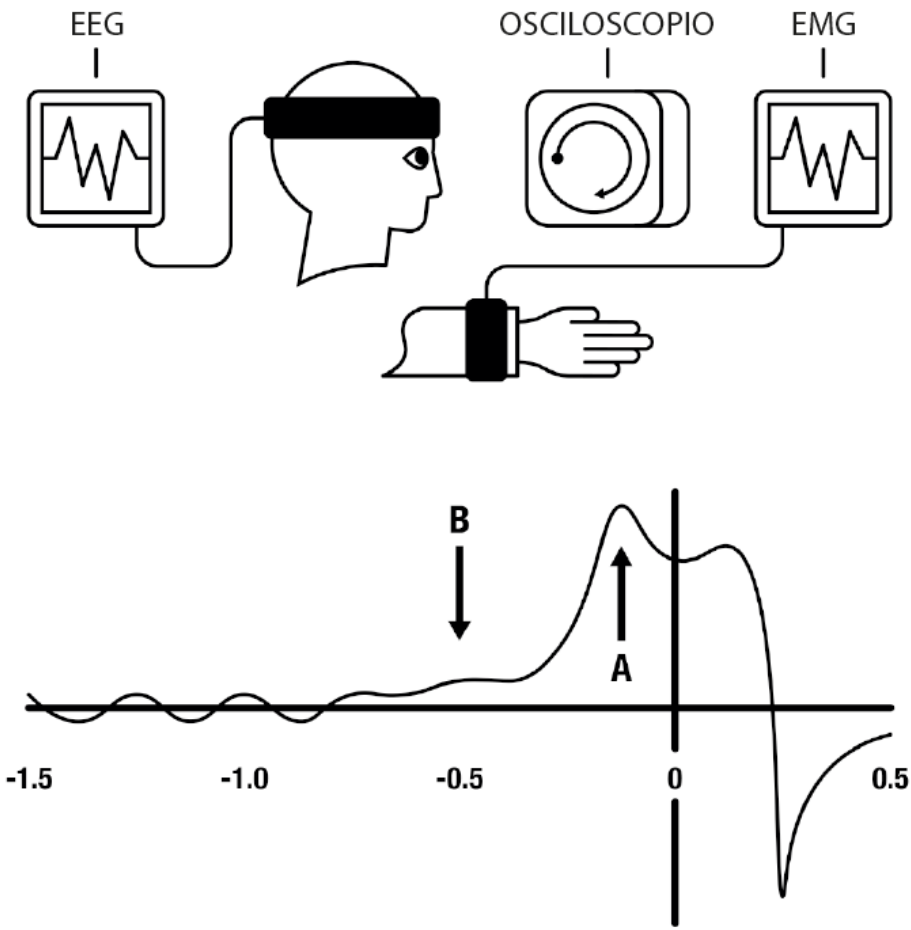

Fig. 19. El famoso experimento de Benjamin Libet sobre la volición [*]

Podemos aclarar ese razonamiento con una analogía. En el juego del martillo de fuerza, una típica atracción de feria, los jugadores golpean con una maza todo lo fuerte que pueden y envían así un pequeño disco de *hockey* lo más alto posible con el objeto de hacer sonar una campana situada en la cima de una torre. Si golpean con la suficiente fuerza, suena la campana; si no, el disco vuelve a caer en silencio. Si una

científica de las ferias examinara las trayectorias de los discos solo en aquellas ocasiones en que los jugadores hicieran sonar la campana, podría concluir erróneamente que toda trayectoria ascendente del disco (el potencial de disposición) provoca siempre un sonido de la campana (la acción voluntaria). Pero, para entender cómo funciona realmente el juego del martillo, la investigadora también tendría que examinar las trayectorias del disco en aquellas otras ocasiones en que la campana *no* suena.

Schurger abordó este problema introduciendo una ingeniosa modificación en el diseño de Libet: los participantes seguirían realizando acciones voluntarias espontáneas, pero oirían de vez en cuando un pitido muy alto que les instaría a realizar esa misma acción de forma no voluntaria, sino forzada por un estímulo. Su descubrimiento clave fue que, en aquellas ocasiones en que sus voluntarios habían reaccionado rápido al pitido, sus EEG mostraron algo muy parecido a un potencial de disposición cuyo inicio, además, se remontaba a bastante antes del pitido en sí, aun cuando no hubieran estado preparándose para ninguna acción voluntaria en esos momentos. Por el contrario, cuando se examinaban las EEG que precedían a las respuestas lentas al pitido, apenas si había señal alguna de nada remotamente parecido a un potencial de disposición.[11]

Schurger interpretó sus datos y postuló que el potencial de disposición *no* es un marcador distintivo del momento en que el cerebro inicia una acción voluntaria, sino un patrón fluctuante de la actividad cerebral que, en ocasiones, supera un umbral y, cuando lo hace, desencadena una acción. Por eso, en el experimento estándar de Libet, se ve en el EEG una pendiente en lento ascenso cuando se mira retrospectivamente en el tiempo desde los momentos en que se produjeron acciones voluntarias. Y por eso también, cuando lo que provoca la acción es un pitido, la respuesta conductual será más rápida si da la casualidad de que esa actividad fluctuante está más próxima al umbral, y será más lenta si resulta estar más lejos del mismo. Esto significa, a su vez, que veremos

algo parecido a un potencial de disposición si miramos hacia atrás desde un momento de reacción rápida (cuando la actividad se situaba casualmente próxima al umbral), pero no cuando lo hagamos desde un momento de reacción lenta (cuando la actividad estaba lejos del umbral).

El elegante experimento de Schurger explica por qué vemos potenciales de disposición cuando buscamos los marcadores neurales distintivos de las acciones voluntarias, y por qué es un error concebirlos como causas específicas de esas acciones. Pero, entonces, ¿cómo deberíamos interpretar esos patrones fluctuantes de la actividad cerebral? Mi interpretación preferida retoma la idea con la que comencé: las experiencias de volición son formas de percepción relacionada con el yo. A través de la óptica del experimento de Schurger, los potenciales de disposición pasan a parecerse mucho a la actividad asociada a la acumulación cerebral de datos sensoriales para elaborar una mejor conjetura bayesiana. Dicho de otro modo, son las huellas dactilares neurales de un tipo especial de alucinación controlada. Acabo de prepararme una taza de té.

Usemos este ejemplo para desarrollar la idea de las experiencias de volición –y también las acciones voluntarias– entendidas como percepciones relacionadas con el yo. Tres son los estados definitorios que caracterizan la mayoría (si no la totalidad) de las experiencias de volición.

El primer rasgo definitorio es la sensación de que *estoy haciendo lo que quiero hacer*. Para un inglés (o inglés a medias) como yo, preparar té es una actividad en perfecta consonancia con mis creencias, valores y deseos psicológicos, así como con mi estado fisiológico dado el momento y dadas las oportunidades (u ofrecimientos) de mi entorno. Tenía sed y había té en la despensa; nadie me impedía ir a buscarlo ni nadie me estaba haciendo tomar un chocolate caliente a la fuerza, así que me preparé un poco de té y lo bebí. (Obviamente, si me fuerzan a hacer algo «contra mi voluntad», puede que aún sienta que mis acciones son

voluntarias a cierto nivel, aunque sean involuntarias a otro).[12]

Si bien preparar té es algo plenamente congruente con mis creencias, valores y deseos, yo no elegí tener esas creencias ni esos valores ni esos deseos. Las acciones voluntarias lo son no porque desciendan de los designios de un alma inmaterial ni porque asciendan de los caprichos de una sopa cuántica. Son voluntarias porque expresan lo que yo, como persona, quiero hacer, aun cuando no pueda elegir esas querencias. Como el filósofo decimonónico Arthur Schopenhauer escribió en su día, «un hombre puede hacer lo que quiere, pero no puede querer lo que quiere».[13]

El segundo rasgo definitorio es la sensación de que *podría haber hecho otra cosa distinta.* Cuando experimento una acción como un acto voluntario mío, lo que caracteriza a esa experiencia no es solo que yo hice X, sino también que hice X y no Y, aun cuando podría haber hecho Y.

Preparé té. ¿Podría haber hecho otra cosa en vez de eso? En cierto sentido, sí. También hay café en la cocina, así que podría haber preparado café. Y cuando estaba haciendo el té, sin duda *tuve la impresión* de que podría haber optado por hacer café en vez de té. Pero no quería café; quería té. Y como no puedo elegir mis querencias, hice té. Dado el estado concreto del universo en aquel momento –en el que se incluye también el estado de mi cuerpo y de mi cerebro–, con sus causas previas (deterministas o no) que se remontan hasta mis inicios como persona semiinglesa bebedora de té, e incluso antes, *yo no podría haber hecho otra cosa.* No puedes volver a reproducir la misma cinta y esperar que produzca un resultado diferente, salvo por unas mínimas diferencias irrelevantes atribuibles al azar. La fenomenología relevante –la *sensación* de que podría haber hecho otra cosa– no es una ventana transparente a través de la que podemos ver cómo funciona la causalidad en el mundo físico.

El tercer rasgo definitorio es que las acciones voluntarias *me parecen salidas de dentro,* en vez de impuestas desde fuera. Esta es la diferencia

entre la experiencia de un acto reflejo (como cuando retiro rápidamente el pie si de pronto me golpeo el dedo gordo, por ejemplo) y la de su equivalente voluntario (como cuando doblo deliberadamente la pierna hacia atrás al disponerme a patear un balón). Es la sensación que tuvo Briony Tallis cuando intentaba pillarse a sí misma su intención consciente de doblar el dedo.

En general, una persona percibe una acción como voluntaria –como fruto de nuestra «libre voluntad»– cuando infiere que sus causas provienen predominantemente de su interior, es decir, que la acción está en consonancia con sus creencias y objetivos, y desconectada de otras causas potenciales alternativas tanto en su cuerpo como en el mundo, y eso le indica la posibilidad de que habría podido actuar de otro modo si así lo hubiera querido. Así es como se sienten las experiencias de volición desde dentro, y así también es como se ven las acciones voluntarias desde fuera.*

El siguiente paso es preguntarnos cómo posibilita y pone en práctica el cerebro tales acciones. Aquí es donde entra en juego el concepto de «grados de libertad» (que es el título del presente capítulo).[14] En ingeniería y matemáticas, un sistema posee tantos grados de libertad como formas múltiples tenga de responder a una situación dada. Básicamente, una roca no tiene ningún grado de libertad, mientras que un tren que circula por una vía única posee solamente uno (ir hacia delante o hacia atrás). Una hormiga puede tener bastantes grados de libertad en lo relativo a cómo su sistema de control biológico responde a su entorno, mientras que tú y yo tenemos infinitamente más grados de libertad aún gracias a la espectacular complejidad de nuestros cuerpos y cerebros.

El comportamiento voluntario depende de que se disponga de la competencia para controlar todos esos grados de libertad de un modo que concuerde con nuestras creencias, valores y objetivos, y esté adaptativamente desconectado de las exigencias inmediatas del entorno y del cuerpo.[15] El cerebro es el encargado de aplicar esta competencia

para controlar, y no desde ninguna región particular donde supuestamente reside la «volición», sino a través de una red de procesos repartidos por muchas regiones del encéfalo. La ejecución de la más simple acción voluntaria –apretar un botón para encender la tetera, o doblar un dedo en el caso de Briony– se apoya en toda esa red. Podemos entender, tal como hace el neurocientífico Patrick Haggard, que esa red pone en práctica tres procesos: un primer proceso de «qué hacer», que especifica qué acción realizar; un proceso subsiguiente de «cuándo hacerlo», que determina el momento de la acción, y un proceso (ya muy hacia el final) de «si hacerlo o no», que posibilita una cancelación o inhibición de última hora.[16]

El componente del «qué hacer» de la volición integra unos conjuntos de creencias, objetivos y valores organizados jerárquicamente, por un lado, y unas percepciones del entorno, por el otro, con el fin de concretar una sola acción entre múltiples posibilidades. Muevo la mano hacia la tetera porque tengo sed, me gusta el té, es el momento ideal del día para beberlo, tengo la tetera a mano, no queda vino, etcétera. Estas percepciones, creencias y objetivos que encajan unos dentro de otros implican a muchas regiones cerebrales diferentes, aunque se concentran especialmente en las partes más frontales del córtex. Por su parte, el componente del «cuándo hacerlo» especifica el momento de una acción seleccionada y está asociado muy íntimamente al impulso subjetivo del movimiento (ese impulso sobre el que se preguntaba Briony Tallis y que Benjamin Libet trató de medir). La base cerebral de este proceso se localiza en las mismas regiones que se asocian al potencial de disposición. De hecho, una suave estimulación cerebral eléctrica de esas regiones –y, en particular, del área motora suplementaria– puede generar un impulso subjetivo a mover(se), aun en ausencia de movimiento alguno.[17] Y, finalmente, el componente del «si hacerlo o no» proporciona un mecanismo de control de último momento sobre si la acción planeada debe proceder o no. Cuando cancelamos una acción en el último instante –en mi caso, tal vez porque se me ha acabado la

leche–, es ese proceso de «inhibición intencional» el que interviene. Estos procesos inhibitorios también son localizables en partes más frontales del cerebro.[18]

Estos procesos entrelazados se desarrollan en un bucle continuo que abarca al cerebro, el cuerpo y el entorno, sin principio ni final, y que pone en práctica una forma continuada y muy flexible de comportamiento orientado a objetivos. Esta red de procesos canaliza un amplio surtido de causas potenciales hacia un único flujo de acciones (o, según el momento, inacciones) voluntarias. Y es sobre la percepción de esa red en funcionamiento (con su proceder circular a través del organismo, hacia fuera –hacia el mundo– y de vuelta al cuerpo de nuevo) sobre la que se sustentan las experiencias subjetivas de la volición.

Además, dado que la acción en sí es una forma de inferencia perceptiva que se cumple por sí misma, como vimos en el capítulo 5, tanto las experiencias perceptuales de la volición como la capacidad de controlar múltiples grados de libertad son dos caras de la misma moneda, es decir, de la misma máquina de predicción. La experiencia perceptiva de la volición es una predicción perceptual que se cumple por sí misma, lo que no deja de ser otra forma característica de alucinación controlada (y tal vez, también *controladora)*.

Existe una razón más por la que experimentamos las acciones voluntarias del modo en que lo hacemos, y es una razón que pone más tierra aún de por medio entre la volición como inferencia perceptiva y la volición como magia dualista. Las experiencias de volición son útiles para guiar la conducta *futura,* tanto como lo son para guiar el comportamiento *presente.*

Como ya hemos visto, la conducta voluntaria es muy flexible. Disponer de la competencia para controlar un gran número de grados de libertad significa que, si una acción voluntaria particular me sale mal, la próxima vez que surja una situación similar, podré probar a hacer algo distinto. Si el lunes intento tomar un atajo con el coche camino del

trabajo pero llego tarde porque me he perdido, el martes tal vez opte por una ruta más larga pero también más fiable. Las experiencias de volición nos ponen de relieve ejemplos de comportamiento voluntario para que podamos prestar atención a sus consecuencias y ajustar conductas futuras para alcanzar mejor nuestros objetivos.

Ya he mencionado antes que nuestro sentido del libre albedrío tiene mucho que ver con la sensación de que «podríamos haberlo hecho de otro modo». Este aspecto factual de la experiencia de la volición es de particular importancia para la función de orientación al futuro que esta tiene. La sensación de que yo podría haber actuado distinto *no* significa que yo pudiera haber hecho otra cosa realmente. La fenomenología de las posibilidades alternativas es útil, más bien, porque nos dice que, en una situación futura similar (aunque no idéntica), realmente podríamos hacer las cosas de manera diferente. Si todas las circunstancias del martes fuesen efectivamente idénticas a las del lunes, entonces yo no podría hacer nada distinto ese día de como lo hice el anterior. Pero eso nunca ocurre. El mundo físico no se duplica con exactitud de una jornada para la siguiente, ni siquiera de un milisegundo para otro. Como mínimo, las circunstancias de mi cerebro ya habrán cambiado, porque habré tenido una experiencia de volición el lunes y me habré fijado en sus consecuencias. Esto, por sí solo, ya es suficiente para afectar al modo en que mi cerebro puede controlar mis muchos grados de libertad cuando salga de casa hacia el trabajo el martes.* La utilidad de sentir que tú «podrías haber hecho las cosas de otro modo» es que, la próxima vez, tal vez las hagas.

¿Y quién es ese «tú»? El «tú» en cuestión es la colección de creencias *a priori* relativas al yo, de valores, de objetivos, de recuerdos y de mejores conjeturas perceptivas que, sumados, componen la experiencia de *ser tú*. Y ahora ya podemos ver que las propias experiencias de volición pueden ser consideradas una parte esencial de ese paquete de cosas que forman la yoidad: son una especie más de alucinación controlada (o controladora) relacionada con el yo. En general, la facultad de ejercer y

experimentar el «libre albedrío» es la capacidad de realizar acciones, hacer elecciones y hasta discurrir pensamientos que son exclusivamente los tuyos propios.

Entonces, ¿el libre albedrío solo es una ilusión? A menudo oímos doctas declaraciones a favor de esa tesis. El renombrado psicólogo Daniel Wegner plasmó muy bien esa visión en su libro *The Illusion of Conscious Will,* que ha tenido una gran influencia desde que se publicó hace casi veinte años.[19] Pero la respuesta correcta a la pregunta es, claro está, que «depende».

Desde luego, el libre albedrío espectral no es real. De hecho, puede que ni siquiera se lo pueda considerar ilusorio. Si la examinamos más de cerca, como aquí hemos hecho, la fenomenología de la volición no consiste tanto en un conjunto de causas inmateriales sin causa como en una alucinación controladora que se cumple por sí misma y se relaciona con unos tipos específicos de acciones: las acciones que parecen venir de nuestro interior. Visto así, el libre albedrío espectral no pasa de ser una solución incoherente a un problema inexistente.[20]

Y aunque, en este capítulo, me he concentrado en ejemplos en los que las acciones voluntarias se acompañan de experiencias de volición claramente sentidas como tales, no siempre es ese el caso. Cuando toco el piano o preparo una taza de té, esas acciones voluntarias se desarrollan la mayor parte del tiempo con un automatismo y una fluidez que resta fuerza no solo a la intuición de que *yo* estoy causando esas acciones de algún modo, sino también a la idea no menos intuitiva –aunque sí examinada con menor frecuencia– de que tales acciones parecen estar causadas por algo. Cuando las personas dicen que están «absortas en lo que están haciendo» o «como en trance» –es decir, profundamente sumidas en una actividad en la que tienen una dilatada práctica–, es muy posible que la fenomenología de la volición esté ausente por completo.[21] Durante gran parte del tiempo, nuestras acciones voluntarias y nuestros pensamientos simplemente «ocurren sin más». Cuando hablamos del libre albedrío, no importa solamente que lo

que las cosas parecen no sea cómo son realmente, pues «lo que las cosas *parecen»* merece un examen más detenido también.

Visto desde otra perspectiva, el libre albedrío no tiene nada de ilusorio. Si tenemos cerebros relativamente indemnes y nos hemos criado en condiciones más o menos normales, cada uno de nosotros posee una capacidad muy real de ejecutar e inhibir acciones voluntarias merced a la facultad de nuestro cerebro para controlar nuestros múltiples grados de libertad. Esta clase de libertad es tanto una libertad *frente a* y una libertad *de* o *para.* Es una libertad *frente a* las causas inmediatas que operan en el mundo y en el cuerpo, y frente a la coacción de las autoridades varias, los hipnotizadores o quienes nos instan a hacer cosas desde las redes sociales. No hablamos, sin embargo, de una libertad que nos libere de las leyes de la naturaleza ni del tejido causal del universo. Por otro lado, también es una libertad *para* actuar conforme a nuestras creencias, valores y objetivos; es la libertad *de* hacer lo que deseemos hacer, y de realizar elecciones con arreglo a quienes somos.

El carácter real de esta clase de libre albedrío queda subrayado por el hecho de que no es algo que podamos dar por descontado. Las lesiones cerebrales o la mala fortuna en la lotería genética o en cuanto al entorno en el que nos ha tocado vivir pueden debilitar nuestra capacidad para ejercer comportamientos voluntarios.[22] Las personas con síndrome de la mano extraña (o «anárquica») realizan acciones voluntarias que no experimentan como si fueran suyas propias, mientras que las que padecen mutismo acinético son incapaces de ejecutar acciones voluntarias de ningún tipo. Un tumor cerebral situado en un sitio complicado puede transformar a un estudiante de Ingeniería en un tirador que asesina en masa a víctimas inocentes en un centro educativo, como le ocurrió a Charles Whitman, el «francotirador de la torre de la Universidad de Texas», o generar en un maestro intachable hasta ese momento una pedofilia galopante (una tendencia que desapareció en cuanto se le extirpó el tumor, y que regresó cuando el cáncer se le

reprodujo en el mismo lugar).[23]

También son muy reales los dilemas éticos y legales planteados por casos como estos. Charles Whitman no eligió desarrollar un tumor cerebral que le oprimiera la amígdala; ¿debía hacérsele responsable de sus actos, entonces? La intuición podría inducirnos a pensar que no, pero cuanto más sabemos acerca de la base cerebral de la volición, más inclinados nos sentimos tal vez a preguntarnos si, para todos y cada uno de nosotros, no se reducirá todo también a una cuestión de «tumores cerebrales».[*] Es un argumento que funciona asimismo a la inversa. Einstein afirmó en una entrevista en 1929 que, como no creía en el libre albedrío, no asumía mérito alguno por nada.[24]

Es igualmente un error calificar de ilusoria la *experiencia de la volición.* Estas experiencias son mejores conjeturas perceptivas, tan reales como cualquier otro tipo de percepción consciente, ya sea del mundo o del yo. Una intención consciente es tan real como una experiencia visual del color. Ninguna de las dos se corresponde directamente con una propiedad definida del mundo (no hay un «rojo real» ni un «azul real» ahí fuera, igual que tampoco hay un libre albedrío espectral), pero ambas contribuyen de manera importante a guiar nuestro comportamiento, y ambas están restringidas por creencias *a priori* y por datos sensoriales. Si las experiencias del color conforman elementos y características del mundo que nos rodea, las experiencias de volición tienen como contenido metafísicamente subversivo la noción de que el «yo» tiene una influencia causal en el mundo. Proyectamos un poder causal en nuestras experiencias de volición del mismo modo que proyectamos rojez en nuestra percepciones de ciertas superficies. Y, ahondando una vez más en la idea de Wittgenstein, saber que esta proyección tiene lugar lo cambia todo y, al mismo tiempo, lo deja todo exactamente igual.

Las experiencias de volición son no solo reales, sino también indispensables para nuestra supervivencia. Son inferencias perceptivas que se cumplen por sí mismas y generan acciones voluntarias. Sin tales

experiencias, no seríamos capaces de manejarnos por los complejos entornos en los que los seres humanos tendemos a florecer y prosperar, ni tampoco podríamos aprender de actos voluntarios previos para hacerlo mejor la siguiente ocasión.[25]

Briony Tallis pensaba que, si pudiera identificar la cresta de la ola de la volición cuando está a punto de romper, podría encontrarse a sí misma. El yo en cuestión es obviamente un yo humano, y bien es cierto que parece haber algo característicamente humano en nuestra capacidad para lidiar con entornos complejos y variables por medio de comportamientos flexibles y voluntarios. No obstante, la facultad de ejercer el libre albedrío podría ser una cuestión de grado, no solo entre nosotros, los diferentes seres humanos, sino también, en general, entre los diversos animales con los que compartimos nuestro mundo.[26]

Y si hacemos extensiva a otras especies la capacidad de ejercer el libre albedrío, ¿qué podríamos decir sobre la extensión de la conciencia en sí?

Ya es hora de que miremos más allá de lo humano.

IV. LOS OTROS

12. MÁS ALLÁ DE LO HUMANO

Desde principios del siglo IX hasta mediados del XVIII no era raro que los tribunales eclesiásticos europeos hallaran a algunos animales penalmente responsables de sus actos. Como resultado, no pocos cerdos fueron ejecutados o quemados vivos, y la misma (mala) suerte corrieron también toros, caballos, anguilas, perros y, al menos en una ocasión (que se sepa), un delfín. De todos modos, en los casi doscientos años documentados por E. P. Evans en la historia del enjuiciamiento penal de los animales que escribió en 1906, los cerdos fueron los delincuentes más habituales, probablemente porque deambulaban con bastante libertad por los pueblos medievales.[1] Acumulaban un variado historial de crímenes: desde comerse a niños hasta devorar hostias consagradas. A veces, se los acusaba de incitar (con sus gruñidos y resoplidos) los crímenes de otro; a menudo, se los ahorcaba, y también se los absolvía de vez en cuando.

Fig. 20. Una cerda y sus lechones son juzgados por el asesinato de un niño Fotografía del Universal History Archive

No resultaba tan fácil acometer judicialmente las plagas (de roedores, langostas, gorgojos y otros animales pequeños o diminutos). En una célebre causa del siglo XVI, el abogado francés Barthelemy de Chassenée logró la exoneración de unas ratas con el inteligente argumento de que no se podía esperar razonablemente que comparecieran en el juicio dados los peligros que para ellas representaban los numerosos gatos que las estarían acechando en el recorrido hasta la sala de la vista. En otros casos, como en varias infestaciones de gorgojos, se llegaron a cursar órdenes escritas a los animales infractores para que abandonaran una propiedad o un cultivo de cebada, a menudo indicando una fecha e incluso una hora específicas.

Por estrambótico que esto pueda parecernos desde nuestra mentalidad del siglo XXI, aquella perspectiva medieval sobre las mentes animales presagiaba en cierto modo el reciente repunte del interés por la conciencia animal y por la cuestión de si la «condición de persona»

puede hacerse extensiva más allá del ámbito humano.[*] La idea de que los animales pudieran entender e incluso someterse razonablemente a los arcanos procedimientos del derecho eclesiástico era (y es) rayana en lo demente. Pero en paralelo a la misma, corría cierto reconocimiento del hecho de que los animales pudieran tener experiencias conscientes y estar dotados de mentes capaces –en cierto sentido– de tomar decisiones. Este reconocimiento de unas mentes conscientes más allá del ámbito humano contrasta drásticamente con la versión cartesiana de la teoría del animal-máquina, en la que los animales carecen del estatus consciente asociado a las mentes racionales. Muchos medievales atribuían sin duda una condición bestial a los animales, pero, para ellos, no eran los meros «autómatas» a los que los reducía el dualismo cartesiano;[2] tenían sus propios mundos interiores, también.

En la actualidad, resultaría extraño y casi perverso, incluso, defender que solo los humanos somos conscientes. Pero ¿qué podemos decir realmente acerca de hasta dónde se extiende el círculo de la conciencia, y acerca de lo diferentes que podrían ser los mundos interiores de otros animales?

Lo primero que hay que decir es que no podemos juzgar si un animal es consciente o no por su capacidad –o incapacidad– para decirnos que lo es. La ausencia de lenguaje no es ninguna prueba de la ausencia de conciencia. Tampoco lo es la ausencia de las llamadas facultades cognitivas de «alto nivel», como la metacognición (que, a grandes trazos, es la capacidad de reflexionar sobre los pensamientos y percepciones propios).[3]

La conciencia animal, allí donde exista, será diferente –y, en algunos casos, muy diferente– de la nuestra. Aunque algunos experimentos con animales pueden arrojar luz sobre los mecanismos de la conciencia en los seres humanos, no es sensato inferir la existencia de conciencia animal basándose solamente en cierta similitud superficial con la del *Homo sapiens*. Si lo hacemos, corremos el doble riesgo de caer en el *antropomorfismo* (la atribución de cualidades humanas a lo que no es

humano) y en el *antropocentrismo* (la tendencia a interpretar el mundo en términos de los valores y las experiencias humanas). El antropomorfismo nos impulsa a ver conciencia parecida a la humana allí donde tal vez no la haya (como cuando creemos que nuestro perro comprende realmente lo que estamos pensando). El antropocentrismo, por su parte, nos cierra los ojos a la diversidad de mentes animales y nos impide reconocer fenómenos de conciencia disímil de la humana allí donde tal vez esté presente (una ceguera de la que es buen ejemplo la concepción cartesiana de los animales-máquina).[4]

Pero, sobre todo, deberíamos ser muy reticentes a asociar en exceso la *conciencia* a la *inteligencia.* La conciencia y la inteligencia no son lo mismo. Cuando se usa la segunda como prueba indiciaria de la primera, se cometen una serie de errores. Se cae en el antropocentrismo: los seres humanos somos inteligentes y conscientes, por consiguiente, para que el animal X sea consciente también debe ser inteligente. Se cae asimismo en el antropomorfismo: vemos inteligencia de apariencia humana en el animal X, pero no en el animal Y, por consiguiente, el animal X es consciente, pero no así el Y. Y fomenta cierta pereza metodológica, pues justifica que se acepten unas pocas capacidades «inteligentes», como el lenguaje y la metacognición (de más fácil acceso ambas que la conciencia propiamente dicha), como suficientes para inferir la presencia de conciencia.

No es que la inteligencia sea irrelevante para la cuestión de la conciencia. A igualdad de condiciones, la inteligencia abre nuevas posibilidades para la experiencia consciente. Puedes estar triste o desilusionado sin tener apenas competencia cognitiva alguna, pero para sentir arrepentimiento –o arrepentimiento anticipatorio– hace falta una capacidad mental suficiente para considerar resultados y cursos de acción alternativos. Hasta las ratas, según sugiere un estudio, podrían experimentar una versión roedora de arrepentimiento, más allá de la mera decepción, cuando las cosas no salen como esperan.[5] *

Las inferencias acerca de la conciencia no humana deben mantener un

delicado equilibrio. Debemos abstenernos de imponer nuestra visión antropocéntrica, pero, al mismo tiempo, no disponemos de muchas más opciones que las que nos dan los propios seres humanos como cantidad conocida con la que comparar resultados y, por lo tanto, como base firme desde la que tratar de llegar más allá, a lo no humano. A fin de cuentas, sabemos que somos seres conscientes y cada vez conocemos mejor el cerebro y los mecanismos corporales implicados en la conciencia humana, algo que podemos usar como base para la extrapolación.

Desde la teoría del animal-máquina desarrollada en este libro, se defiende que la conciencia es algo más estrechamente conectado con el hecho de estar *vivo* que con el de ser *inteligente.* Como es natural, esto es tan válido para los otros animales como lo es para nosotros, los humanos. Desde ese punto de vista, la conciencia podría estar más extendida de lo que podría parecer si tomáramos la inteligencia como criterio prioritario. Pero no significa que donde haya vida habrá también conciencia.[6]

Buscar la conciencia más allá de lo humano es como aventurarse por un lago helado desde una orilla desoladoramente glacial. Hay que ir dando un paso cauteloso tras otro, siempre probando la sensación de solidez del hielo bajo nuestros pies.

Comencemos por los mamíferos, un colectivo en el que están incluidos las ratas, los murciélagos, los monos, los manatíes, los leones, los hipopótamos y, por supuesto, también los humanos. Yo creo que *todos los mamíferos son conscientes.* Por supuesto, no lo sé con certeza, pero estoy bastante seguro de ello. Esta afirmación no se basa en ninguna similitud superficial entre esos animales y los seres humanos, sino en el hecho de que poseen mecanismos compartidos. Dejando a un lado el tamaño absoluto del cerebro (que tiene más que ver con el tamaño del cuerpo que con ninguna otra cosa), los cerebros de los mamíferos son sorprendentemente similares entre especies.[7]

En 2005, el científico cognitivo Bernard Baars, David Edelman –hijo

de Gerald Edelman y experto en cognición animal– y yo mismo elaboramos una lista de propiedades de la conciencia humana que considerábamos que podían ser fácilmente testables en otros mamíferos. Incluimos un total de diecisiete propiedades distintas.[8] Hasta cierto punto, se trataba de una cifra arbitraria, pero venía a mostrar lo razonable que resultaba formular hipótesis experimentalmente contrastables acerca de la conciencia animal.

Las primeras propiedades en las que pensamos tenían que ver con características anatómicas del cerebro. En términos de «cableado» cerebral, los principales elementos neuroanatómicos que se asocian con fuerza a la conciencia humana están presentes en todas las especies de mamíferos. Hay un córtex de seis capas, un tálamo muy interconectado con esa corteza, un tronco del encéfalo en una posición profunda y toda una serie de otras características compartidas –como, por ejemplo, los sistemas de neurotransmisores– que aparecen sistemáticamente implicadas (en los seres humanos, al menos) en el flujo instante a instante de la experiencia consciente.

También hay rasgos comunes en lo relativo a la actividad cerebral. Entre los más llamativos están las variaciones en la dinámica del cerebro entre el momento en que los animales se duermen y el momento en que despiertan, es decir, en la dinámica subyacente al *nivel* consciente. En estados de vigilia normales, todos los mamíferos muestran una actividad cerebral eléctrica irregular, de baja amplitud y rápida. Y cuando les llega el sueño, todos los cerebros de mamíferos cambian a una dinámica cerebral más regular y de gran amplitud. Estos patrones y cambios se asemejan mucho a lo observado en humanos durante la vigilia y el sueño. También la anestesia general surte efectos similares en todas las especies de mamíferos: una interrupción casi omnímoda de la comunicación entre regiones cerebrales que se acompaña de una absoluta ausencia de respuesta conductual.[9]

Existen, desde luego, múltiples diferencias también, sobre todo en lo referente a los patrones del sueño. Las focas y los delfines duermen con

medio cerebro despierto en todo momento; los koalas duermen unas veintidós horas al día, y las jirafas se las arreglan con menos de cuatro, mientras que las orcas recién nacidas no duermen para nada durante su primer mes de vida.[10] Casi todos los mamíferos presentan fases de sueño de movimientos oculares rápidos (REM o MOR), pero las focas solo las tienen cuando duermen en tierra, y los delfines no parecen tenerlas nunca.

Por otra parte, entre unas especies de mamíferos y otras tiene que haber desemejanzas también en el *contenido* consciente, además de en el nivel consciente. Buena parte de esa variación puede atribuirse a diferencias en los tipos de percepción dominantes. Los ratones dependen sobre todo de sus bigotes; los murciélagos, de su sonar ecolocalizador, y las ratas topo desnudas, de su agudo sentido del olfato, sobre todo, cuando se encuentran con otros individuos de su especie. Estas diferencias en cuanto a dominancia perceptual sin duda se traducen en que cada animal habite en un mundo interior diferenciado de los de las otras especies.[11] *

Más intrigantes aún son aquellas diferencias relacionadas con la manera de experimentar la *yoidad*. En los seres humanos, un notable marcador del desarrollo de una autoconciencia de alto nivel –como la que relacionamos con la identidad personal– es la facultad de reconocerse a uno mismo en un espejo. En las personas, esta capacidad de «autorreconocimiento en el espejo» tiende a desarrollarse entre los dieciocho y los veinticuatro meses de edad. Eso no significa que los niños más pequeños aún carezcan de conciencia, sino solamente que puede que la consciencia de sí mismos como individuos separados de otros no esté plenamente formada todavía antes de esa edad.

En animales, se ha investigado abundantemente la capacidad de autorreconocimiento aplicando un test desarrollado por el psicólogo Gordon Gallup Jr. en la década de 1970.[12] En la versión clásica de su prueba de autorreconocimiento en el espejo, primero se anestesia al animal y luego se le practica una marca –por lo general, con pintura o

un adhesivo– en un lugar de su cuerpo que normalmente no pueda ver por sí mismo. Luego, se deja que el animal, una vez se le hayan pasado los efectos de la anestesia, interactúe con un espejo para que él mismo pueda verse la marca. Si, tras haber mirado hacia el espejo, se busca espontáneamente la marca en su propio cuerpo en vez de investigar la imagen especular que está viendo, se puede considerar que ha superado la prueba. La lógica en la que se basa ese criterio es que, si actúa así, la criatura ha reconocido que la imagen del espejo es un reflejo de su propio cuerpo y no el cuerpo de otro animal.

¿Quiénes superan la prueba del espejo? Entre los mamíferos, algunos grandes simios,[13] unos cuantos delfines y orcas, y un único elefante eurasiático. Un largo desfile de otras criaturas del grupo de los mamíferos, entre las que se incluyen pandas, perros y monos varios, la han suspendido… Al menos, hasta el momento. Dado lo intuitivo que es el autorreconocimiento en el espejo para nosotros, los seres humanos, y lo cognitivamente competentes que, por lo demás, parecen ser muchos de estos mamíferos que no se autorreconocen, llama la atención lo escasa que es esta lista de aprobados. No hay ninguna demostración convincente de que algún no mamífero haya superado la prueba del espejo, aunque las mantarrayas y las urracas podrían estar cerca, y actualmente existe cierta controversia sobre si lo estará también el lábrido limpiador.[14]

Ahora bien, es posible que los animales no superen la prueba del espejo por muchas razones además de por carecer de capacidades de autorreconocimiento. Entre esos motivos podrían estar que no les gusten los espejos, que no entiendan cómo funcionan o incluso que simplemente prefieran evitar el contacto visual. Conscientes de ello, los investigadores siempre están desarrollando nuevas versiones del test, astutamente diseñadas para diferentes mundos interiores (es decir, para diferentes mundos perceptivos). Por ejemplo, el autorreconocimiento en perros puede testarse ahora mediante «espejos olfativos», aunque sigue sin irles muy bien en la prueba *(dognición* –por *dog*– es el encantador

nombre con el que se conoce la cognición en perros).[15] Existe la posibilidad de que, a medida que el ingenio experimental avance en su escalada y su despliegue, varias de las especies que están ahora a un lado de la línea pasen a estarlo del otro, del luminoso lado del autorreconocimiento certificado por el espejo. Pero, aun en ese caso, la diversidad de pruebas del espejo existentes –y la incapacidad de muchos animales para superar siquiera tests muy adaptados a su especie– nos da a entender que es muy probable que existan diferencias muy sensibles en cuanto a cómo experimentan los mamíferos el «ser ellos mismos».

Estas diferencias me sorprenden muy especialmente en el caso de los monos. Aunque los chimpancés y los grandes simios son nuestros vecinos evolutivos más cercanos, los monos tampoco son parientes lejanos nuestros y hace tiempo que se usan en experimentos neurocientíficos como «modelos primates» de los seres humanos, sobre todo, en lo relacionado con la visión. En algunos estudios, incluso se ha entrenado a monos para que atestigüen (o, en el lenguaje experimental, «informen»), por ejemplo, accionando una palanca, si han «visto» algo o no.[16] Estos experimentos se pueden comparar directamente con estudios con humanos en los que las personas dicen lo que experimentan (o no experimentan): se obtiene así un equivalente para primates de un método clave en la investigación de la conciencia.[17]

Dadas sus muchas similitudes con los humanos, no me cabe duda alguna de que los monos poseen cierto tipo de yoidad consciente. Si pasas algún tiempo con ellos, no puedes evitar tener la impresión totalmente convencida de que estás entre entes dotados de conciencia (es decir, entre otros *yos* conscientes).

Esa fue mi experiencia en julio de 2017, cuando pasé un día en Cayo Santiago, una pequeña isla situada frente a la costa oriental de Puerto Rico, en el Caribe. A Cayo Santiago también se la conoce como Isla de los Monos porque sus únicos residentes permanentes son macacos Rhesus (más de un millar de ellos). Se trata de una población trasplantada allí desde Calcuta en 1938 por un excéntrico zoólogo

estadounidense, Clarence Ray Carpenter, cansado de tantos largos viajes hasta la India.[18] Mientras recorría a pie Cayo Santiago en aquel cálido día de verano, acompañado por la psicóloga de Yale Laurie Santos (y los miembros de un equipo de rodaje), decenas de monos iban a lo suyo, recelosos y, al mismo tiempo, tolerantes con la presencia de aquellos humanos con su torpe deambular. Dos monos se turnaban para subirse a un árbol y saltar desde sus ramas a un estanque que había justo debajo, y parecían hacerlo sin más razón que el puro placer espontáneo de saltar y zambullirse. Se estaban *divirtiendo*.[*]

No menos convincentes en ese mismo sentido resultan los vídeos de monos capuchinos que reaccionan ante injusticias deliberadas. En uno de ellos, popularizado por el primatólogo Frans de Waal, dos monas alojadas en jaulas adyacentes reciben una recompensa en turnos sucesivos a cambio de pasarle una pequeña piedra a un experimentador.[19] La mona 1 pasa la piedra entre la rejilla de la jaula y se la recompensa con una rodaja pequeña de pepino, que se come encantada. La mona 2 hace lo mismo y no recibe pepino, sino una uva, que es un bocado mucho más apetitoso. La mona 2 se come la uva mientras la mona 1 se la queda mirando. Cuando la mona 1 repite la acción durante su turno y le vuelven a dar pepino, lo mira, se lo arroja de vuelta al experimentador y sacude las rejas de la jaula con gesto de aparente indignación.

La diversión y los berrinches nos inducen poderosamente a intuir algo más. Estos comportamientos son tan distintivos que es casi imposible interpretarlos de otro modo que no sea como la manifestación externa de unos estados interiores aparentemente parecidos a los humanos. Cuando vemos a un mono comportándose así, intuimos la presencia no ya de otro ser consciente, sino de un ser consciente *como nosotros*. Pero el caso es que, como ya he mencionado aquí, los monos han suspendido reiteradamente la prueba del espejo.[20] Así que, si bien los monos son sin duda seres conscientes, y aunque también creo que experimentan cierto tipo de yoidad, no los podemos considerar unas personitas peludas sin

más.

La influencia del antropomorfismo y el antropocentrismo en nuestras intuiciones se hace más evidente aún cuando extendemos nuestra mirada más allá de los mamíferos, y, en especial, cuando miramos a algunos de nuestros parientes evolutivos más lejanos.

En el verano de 2009, David Edelman y yo pasamos una semana con una docena aproximada de pulpos comunes –*Octopus vulgaris*–, una especie de cefalópodos.[*] Estábamos visitando al biólogo Graziano Fiorito, un destacado experto en cognición y neurobiología de esa clase de invertebrados. Aunque ha pasado ya más de una década desde entonces, aquella semana sigue siendo una de las más memorables que recuerdo de todos mis años como científico.

El laboratorio de pulpos de Fiorito –que forma parte del renombrado instituto de investigación italiano Stazione Zoologica– se encuentra en un sótano frío y húmedo que está situado justo debajo de un acuario público, en pleno corazón de Nápoles: un buen rincón fresco en el que refugiarse del ardiente calor veraniego a nivel de calle. La mayor parte de mi semana allí la pasé en compañía de tan fascinantes criaturas, observando cómo cambian de forma, color y textura, y fijándome también en qué es aquello a lo que ellas prestan atención. Un día, mientras trataba de comparar el aspecto siempre cambiante de un pulpo en particular con los dibujos del libro de Fiorito, *A Catalogue of Body Patterning in Cephalopoda,*[21] escuché una especie de «¡plaf!» sordo seguido de una especie de culebreo. Había dejado entreabierta la tapa de un tanque y aquella misma criatura estaba tratando de huir por allí. Sigo convencido de que esperó a que me dejara llevar por una falsa sensación de seguridad y me volviera hacia otro lado durante el tiempo suficiente.

Mientras yo andaba liando las cosas de vez en cuando, David estaba preparando un experimento sobre percepción visual y aprendizaje. Sumergía objetos con diferentes formas en el tanque de agua de un pulpo, y acompañaba algunos de ellos de un sabroso cangrejo. Se

trataba de ver si el pulpo podía aprender a asociar objetos concretos con una recompensa. No recuerdo exactamente cómo salió el estudio, pero sí me acuerdo muy claramente de un episodio en particular.

En el laboratorio de Fiorito, había dispuestas dos filas de tanques de agua flanqueando un pasillo central, con un pulpo en cada tanque. (Los pulpos, por lo general, no son criaturas sociales y pueden tener incluso impulsos caníbales). Aquel día, David había seleccionado un tanque de la fila de la izquierda, aproximadamente hacia la mitad. Cuando me acerqué para ver lo que pasaba en él, me asombré de ver que todos los pulpos del otro lado del pasillo se apretaban contra el vidrio frontal de sus tanques, todos y cada uno mirando fijamente a David mientras sumergía en diversas ocasiones sus objetos en el tanque que había elegido. Los pulpos que lo observaban parecían estar tratando de averiguar qué estaba sucediendo allí sin otra razón que la del simple interés o curiosidad.

Estando entre pulpos, aunque no fuera durante mucho tiempo, tuve la impresión de encontrarme ante una presencia inteligente y consciente, aunque muy distinta de cualquier otra y, desde luego, con una inteligencia y una conciencia muy diferentes de sus manifestaciones humanas. Se trataba, desde luego, de una impresión subjetiva e inevitablemente contaminada por los sesgos propios del antropomorfismo y el antropocentrismo, y por la que se me podía reprochar que tal vez estuviese tomando la inteligencia como señal de sintiencia. Pero, en cualquier caso, los pulpos son objetivamente asombrosos y pasar un tiempo con ellos puede ensanchar mucho nuestra intuición de lo diferente que una conciencia no humana podría ser.

El antepasado común más cercano de humanos y pulpos vivió hace unos seiscientos millones de años. Poco se conoce de esa criatura arcaica. Tal vez fuera alguna especie de platelminto. Fuera como fuere, debió de ser un animal muy simple. Las mentes de los pulpos no son derivados acuáticos de la nuestra, ni, de hecho, de la de ninguna otra especie de vertebrado, ni pasada ni presente. La mente de un pulpo es

un experimento evolutivo independiente que es lo más próximo a la mente de un extraterrestre que probablemente podamos encontrar sobre la faz de este planeta.[22] En palabras del filósofo y buceador Peter Godfrey-Smith, «si queremos comprender *otras* mentes, las de los cefalópodos son las más "otras" de todas».[23]

El cuerpo de un pulpo causa admiración por sí solo. El pulpo común, *Octopus vulgaris,* posee ocho apéndices a modo de brazos, tres corazones que bombean sangre azul, un mecanismo de defensa con tinta y un sistema muy desarrollado de propulsión a chorro. Un pulpo puede cambiar de tamaño, forma, textura y color a voluntad, y puede hacerlo todo al mismo tiempo si es necesario. Es un animal líquido: aparte de un pico huesudo en el centro, el resto del cuerpo del pulpo es blando, lo que le permite escurrirse por rendijas imposiblemente pequeñas (como pude comprobar por mí mismo en la Stazione Zoologica).

Estos cuerpos extraordinarios se complementan con unos sistemas nerviosos muy sofisticados. El *Octopus vulgaris* posee unos quinientos millones de neuronas, aproximadamente seis veces las de un ratón. A diferencia de lo que ocurre en los mamíferos, la mayoría de esas neuronas –unas tres quintas partes– están en sus tentáculos en vez de en su cerebro central (un cerebro que, de todos modos, cuenta con hasta cuarenta lóbulos anatómicamente diferenciados). Otra característica inusual de los pulpos es que sus cerebros carecen de mielina, el material aislante que, en los cerebros de los mamíferos, ayuda a que se desarrollen y funcionen las conexiones neurales de largo alcance.[24] El sistema nervioso del pulpo está, pues, más repartido y menos integrado que los de los mamíferos de tamaño y complejidad parecidos. Por consiguiente, la conciencia de un pulpo –suponiendo que la tenga– podría estar también más diseminada y menos integrada, e incluso carecer de un «centro» unitario.[25]

Los pulpos hacen las cosas de un modo distinto incluso a nivel de sus genes. En la mayoría de organismos, la información genética del ADN se transcribe directamente en secuencias más cortas de ARN (ácido

ribonucleico), que se usan luego para fabricar proteínas, que son los talleres moleculares de la vida. Es un principio consolidado, de manual, de la biología molecular. Sin embargo, en 2017, el descubrimiento de que las secuencias de ARN en los pulpos (y en unos cuantos cefalópodos más) pueden experimentar un grado de edición significativo antes de traducirse en una proteína dio un vuelco a dicho principio. Es como si el pulpo fuera capaz de reescribir partes de su propio genoma sobre la marcha. (Ya se habían detectado casos de edición del ARN anteriormente en otras especies, pero son casos en los que dicha edición tiene un papel relativamente menor). Además, en el caso del pulpo, buena parte de su edición del ARN parece estar relacionada con el sistema nervioso. Algunos investigadores han sugerido que las impresionantes aptitudes cognitivas de los pulpos podrían fundamentarse, en parte, en esa prolífica capacidad suya de reescritura del genoma.[26]

Y vaya si impresionan sus aptitudes cognitivas.[27] Pueden recuperar objetos escondidos –normalmente, sabrosos cangrejos– en el interior de cubos de plexiglás metidos dentro de otros; pueden hallar la salida de laberintos complejos; pueden probar un amplio elenco de acciones distintas hasta resolver un problema concreto, y, como el propio Fiorito ha demostrado en la Stazione, pueden aprender simplemente observando a otros pulpos. Más asombrosos aún son los testimonios anecdóticos de avistamientos del comportamiento de estos animales en estado salvaje. Uno de los más extraordinarios aparece en un episodio de la serie de televisión de la BBC *Planeta azul II,* donde se puede ver a un pulpo, sorprendido en su entorno natural, cubriéndose con conchas y otros fragmentos de detritus del fondo marino para esconderse de un depredador: un tiburón que ronda por aquellas aguas.[28]

Estas proezas de inteligencia cefalópoda son prueba ciertamente convincente de la presencia de una mente en acción. Pero ¿qué clase de mente? Ya he comentado que no deberíamos cargar demasiado las tintas en la inteligencia como marcador de la conciencia. Así que ¿qué

podemos decir sobre el modo particular de ser de los pulpos? Para poder abordar bien esa pregunta, necesitamos conectar el comportamiento de los pulpos con su percepción.

La del camuflaje es posiblemente la habilidad más alienígena del catálogo de capacidades de los cefalópodos. Sin conchas duras que los protejan, su supervivencia depende a menudo de su facilidad para confundirse con el ambiente.[29] Pueden imitar el color, la forma y la textura de su entorno con tal perfección que, tanto para ti como para mí –y para muchos de sus depredadores potenciales–, el animal resultaría prácticamente invisible incluso a uno o dos metros de distancia.

Los pulpos ajustan su color al de su entorno valiéndose de un sistema de *cromatóforos* de exquisita precisión. Son unos pequeños receptáculos elásticos, repartidos por toda la piel, que producen una coloración roja, amarilla o marrón cuando se abren a instancias de órdenes neurales llegadas principalmente desde unos lóbulos cerebrales dedicados específicamente a dichos cromatóforos. El funcionamiento exacto del sistema todavía no se conoce del todo. Parte de la dificultad estriba en que los pulpos deben invisibilizarse no a ojos de otros pulpos, sino a los de los depredadores, que también ven el mundo a sus diferenciados modos respectivos. Su sistema de camuflaje debe de tener codificado de algún modo cierto conocimiento acerca de las facultades visuales de esos depredadores.

Aún más sorprendente resulta el hecho de que los pulpos hacen todo esto a pesar de ser daltónicos. Las células sensibles a la luz en los ojos humanos responden a tres longitudes de onda lumínicas diferentes y, al mezclarlas, crean todo un universo de colores. Las células en los ojos de los pulpos, sin embargo, no contienen más que un fotopigmento. Los pulpos pueden notar la dirección de la polarización de la luz –igual que tú o que yo cuando llevamos gafas de sol polarizadas–, pero no pueden deducir colores a partir de combinaciones de longitudes de onda diferentes. Esa misma ceguera al color es extensiva a las células sensibles a la luz que llevan incrustadas a lo largo y ancho de la piel,

pero, a cambio, los pulpos pueden «ver» con la piel, además de con los ojos. Podemos añadir a todo eso que se cree que el control de los cromatóforos por parte de los pulpos es de «circuito abierto», lo que significa que las neuronas del lóbulo de los cromatóforos no generan ninguna copia interna evidente de las señales que envían a los cromatóforos de la piel. Eso quiere decir que puede que el cerebro central del animal ni siquiera sepa lo que está haciendo su piel.[30]

Cuesta hacerse una idea de lo que esto puede significar en términos de cómo un pulpo experimenta su mundo, y su propio cuerpo dentro de ese mundo. Su propia piel cambia de color de modos y maneras que él mismo no puede ver y de las que ni siquiera envía señales a su cerebro. Y puede que algunas de esas adaptaciones tengan lugar como consecuencia de un control puramente local: el tentáculo recibe sensaciones de su entorno inmediato y cambia su aspecto sin que el cerebro central tenga implicación alguna. El supuesto antropocéntrico de que podemos ver y sentir lo que les sucede a nuestros propios cuerpos no es válido en su caso. No es de extrañar que los pulpos no muestren señal alguna de que vayan a superar nunca la prueba del espejo.

Además de la vista, los pulpos comparten con los mamíferos y otros vertebrados algunas de las otras modalidades sensoriales clásicas. Pueden saborear, oler y tocar; también pueden oír, aunque no muy bien. También ahí presentan particularidades muy peculiares, pues los pulpos pueden saborear tanto con sus ventosas como con partes de su boca central.[31] Esto parece indicarnos de nuevo que estas criaturas seguramente presentan una extraordinaria descentralización mental.

La idea de una conciencia descentralizada es particularmente difícil de concebir para nosotros, sobre todo en lo tocante a experiencias de propiedad corporal. Como vimos en el capítulo 8, en los seres humanos ese aspecto de la yoidad consciente se puede alterar con sorprendente facilidad, simplemente induciendo al cerebro a cambiar sus mejores conjeturas bayesianas acerca de qué forma parte del cuerpo y qué no.

Bastante difícil es ya para nosotros, los humanos, hacer un seguimiento de nuestro propio cuerpo con solo cuatro extremidades que tienen sus movimientos limitados a unas pocas articulaciones. Imagínate lo que puede ser para un pulpo, con sus ocho brazos increíblemente flexibles que se enrollan y se desenrollan en varias direcciones a la vez: el desafío es formidable. E igual que la sensación está parcialmente descentralizada en esos brazos, también lo está el control. Los tentáculos del pulpo se comportan como animales semiautónomos: cuando a un pulpo se le amputa uno de ellos, el miembro extirpado puede seguir ejecutando durante un tiempo secuencias complejas de acciones, como atrapar trozos de comida.

Estos grados de libertad, sumados al control descentralizado, plantean un intimidante reto para cualquier cerebro central que intente mantener una percepción unitaria y unificada de lo que forma parte de su cuerpo y lo que no. Y por eso es posible que los pulpos ni siquiera se preocupen por querer saberlo. Por extraño que parezca, puede que el modo del ser de un pulpo no incluya una experiencia de la propiedad corporal que se parezca en lo más mínimo a la que correspondería al caso de los seres humanos y otros mamíferos.

Eso no quiere decir que los pulpos no distingan su «yo» de «lo otro». Es evidente que lo hacen y que necesitan hacerlo. Para empezar, tienen que evitar enredarse consigo mismos. Las ventosas del tentáculo de un pulpo se adhieren por reacción refleja a casi todos los objetos que pasan por su lado, pero no a otros tentáculos del mismo animal, ni tampoco a la parte central de su cuerpo. Esto demuestra que los pulpos son capaces de algún modo de discernir qué es su cuerpo y qué no.

Curiosamente, esa capacidad depende de un sistema sencillo pero eficaz de autorreconocimiento basado en el gusto. Los pulpos secretan una sustancia química característica por toda su piel. Esa sustancia actúa como señal detectable por las ventosas, por lo que estas no se adhieren por reflejo a nada que esté impregnado de aquella. De este modo, un pulpo puede distinguir *qué* parte del mundo es parte de sí mismo y

cuáles no, aun cuando no sepa necesariamente *dónde* está situado su cuerpo en el espacio. Este descubrimiento se corroboró por medio de una serie de experimentos ciertamente macabros en los que los investigadores ofrecían a unos tentáculos amputados de pulpo otros tentáculos amputados, algunos con la piel original, otros sin ella.[32] Los tentáculos amputados se adherían al momento a los brazos a los que se les había quitado la piel, pero no a aquellos en los que esta se había dejado intacta.[*]

Lo que esto quiere decir a propósito de las experiencias de corporeidad en un pulpo es algo muy difícil de imaginar para unos mamíferos como nosotros. Es posible que el pulpo en su totalidad solo tenga una percepción muy difusa de qué es su cuerpo y dónde está, aunque él probablemente no experimente esa percepción como algo difuso. E incluso es posible que exista algo así como un modo de ser un tentáculo de pulpo.

Los pulpos ponen seriamente a prueba nuestras ideas intuitivas sobre lo diferente que la conciencia animal puede llegar a ser de la nuestra. Pero con ese salto directo de los monos a los cefalópodos, nos hemos dejado un enorme surtido de posibilidades intermedias. Tras las seguras costas de la conciencia mamífera se extienden vastos territorios de potenciales estados de consciencia en animales que van desde los loros hasta protozoos unicelulares como los paramecios. Ahora, con la vista puesta sobre semejante extensión, volvamos sobre la pregunta, más fundamental, de qué animales tienen probablemente algún tipo de experiencia consciente, es decir, de qué animales tienen la «luz encendida», por leve que sea su resplandor.[33]

Las aves nos dan muy buenos motivos para suponerles sintiencia. Los cerebros aviares, pese a ser significativamente diferentes de sus homólogos en los mamíferos, muestran una organización que se corresponde de forma bastante aproximada con la del córtex y el tálamo de estos últimos. Muchas especies de ave llaman también la atención por su inteligencia.[34] Los loros pueden contar, las cacatúas pueden bailar y

las charas californianas pueden almacenar comida en función de sus necesidades futuras. Aunque estos ejemplos de conductas inteligentes dan a entender que algunas aves pueden experimentar estados de conciencia complejos, recuerda que la inteligencia no es ninguna prueba de la presencia de consciencia. Es probable que las aves que no esconden comida ni hablan ni bailan tengan sus propias experiencias conscientes también.[35]

A medida que nos alejamos aún más de los mamíferos, los indicios se van volviendo más escasos y fragmentarios, y las inferencias de la presencia de conciencia, más inciertas y conjeturales.[36] En vez de basar esas inferencias en la similitud con los cerebros y las conductas de los mamíferos, una mejor estrategia consistiría en adoptar la perspectiva del animal-máquina –la mía, no la cartesiana–, que atribuye el origen y la función de la percepción consciente a la regulación fisiológica y a la preservación de la integridad del organismo. En ese caso, un buen sitio donde buscar indicios de la existencia de consciencia podría ser en el modo en que los animales reaccionan a hechos supuestamente dolorosos para ellos.

Esta estrategia no solo es sensata desde el punto de vista científico, sino también justificable por imperativos éticos. Las decisiones sobre el bienestar animal no deberían fundamentarse en la similitud de la criatura en cuestión con los seres humanos ni en si logra superar cierto umbral arbitrario de competencia cognitiva, sino en su capacidad para sentir dolor y sufrir.[37] Y si bien existe una infinidad de formas en que las criaturas vivas pueden sufrir, las más ampliamente compartidas por las diversas especies son aquellas que implican ataques a su integridad fisiológica.

Cuando se han buscado, se han encontrado abundantes pruebas de respuestas adaptativas a sucesos dolorosos en diversas especies animales. La mayoría de vertebrados (animales con espina dorsal) cuidan de las partes corporales que se lesionan. Incluso el diminuto pez cebra está dispuesto a pagar un «precio» por conseguir alivio para el

dolor cuando se lesiona: no tiene reparo en abandonar el entorno natural en que se encuentra para pasar a una pecera desierta y muy iluminada previamente bañada en una sustancia analgésica. Que eso implique que los peces son seres conscientes –y recordemos que hay *muchísimos* tipos de pez– no está ni mucho menos claro, pero es ciertamente indicativo de algo.

¿Y los insectos? Las hormigas no cojean cuando se dañan una pata. Sin embargo, es posible que los exoesqueletos duros que recubren sus cuerpos los hagan menos susceptibles de sentir dolor; además, los cerebros insectiles poseen variantes del sistema de neurotransmisores opiáceos que se asocia generalmente al alivio del dolor en otros animales.[38] En un estudio reciente, se descubrió que, tras una lesión, la mosca de la fruta, *Drosophila melanogaster,* evidencia una hipersensibilidad –a estímulos que antes no le resultaban dolorosos– que recuerda en cierto modo al «dolor crónico» en humanos.[39] Y llama mucho la atención el hecho de que los fármacos anestésicos parecen surtir efecto en *todos* los animales, desde los bichitos unicelulares hasta los primates avanzados.[40]

Todo esto es indiciario, pero no concluyente.

Llega un punto a partir del cual resulta difícil afirmar nada sustancial. Mi intuición –que no pasa de ser eso, una intuición– es que debe de haber animales que no son miembros en absoluto del círculo de la conciencia. Una de las razones por las que pienso esto es que incluso en los mamíferos, con nuestros cerebros complejos y nuestros sistemas perceptivos bien afinados y orientados a la preservación de nuestra integridad fisiológica, continúa siendo relativamente fácil caer en la inconsciencia. La experiencia consciente tiene un papel central en nuestras vidas, pero eso no significa que su base biológica sea simple y directa. Cuando llegamos al nivel de seres como el nematodo, con sus nimias trescientas dos neuronas, me cuesta mucho imaginarles estado consciente de ningún tipo. Y me parece imposible que el unicelular paramecio dé la talla para algo así.

Dejando a un lado sus inevitables incertidumbres, el estudio de la conciencia animal nos brinda dos beneficios de gran calado. El primero es el reconocimiento del hecho de que la manera en que los seres humanos experimentamos el mundo y el yo no es la única posible. Habitamos una angostísima región de un inmenso espectro de mentes conscientes posibles, y la investigación científica que se ha hecho sobre ese espectro hasta el momento equivale a poco más que unos pocos destellos en la oscuridad.[41] El segundo de los beneficios es una recién descubierta humildad. Al contemplar por fin la enorme diversidad de la vida sobre la tierra, tal vez comencemos a valorar más –y a dar menos por descontada– la riqueza de la experiencia subjetiva en toda su variedad y peculiaridad, tanto en nosotros mismos como en otros animales. Y puede que también hallemos motivación renovada para minimizar el sufrimiento comoquiera (y dondequiera) que se presente.

Inicié este capítulo defendiendo que la conciencia y la inteligencia no son lo mismo, y que la conciencia tiene más que ver con el hecho de estar vivos que con el de ser inteligentes. Quisiera acabarlo con una afirmación más contundente aún. No solo puede existir conciencia sin mucha inteligencia que la acompañe –no hace falta ser inteligente para sufrir, por ejemplo–, sino que la inteligencia puede existir sin conciencia también.

La posibilidad de que la inteligencia no vaya necesariamente acompañada de la capacidad de sufrimiento nos lleva a la etapa final de nuestro viaje a través de la ciencia de la conciencia. Llegó la hora de hablar de la inteligencia artificial y de si es posible o no que alguna vez existan máquinas conscientes.

13. MENTES-MÁQUINA

En Praga, a finales del siglo XVI, el rabino Judah Loew ben Bezalel extrajo arcilla de las orillas del río Moldava y, con ella, moldeó una figura de aspecto humano: un golem.[1] Este golem –al que llamaron Josef, o Yoselle– fue creado para la defensa de la grey del rabino frente a los pogromos antisemíticos y, al parecer, fue bastante eficaz. Una vez activados por ensalmo mágico, los golems como Josef se podían mover, tenían consciencia y obedecían. Pero con Josef la cosa terminó por degenerar y su conducta pasó de la ciega obediencia a la violencia monstruosa. Al final, el rabino consiguió revocar su conjuro e, *ipso facto,* su golem se deshizo en pedazos en el propio recinto de la sinagoga. Hay quienes dicen que sus restos continúan ocultos en Praga aún hoy en día, tal vez en una tumba, tal vez en un desván, tal vez aguardando pacientemente a ser reanimados.

El golem del rabino Loew nos recuerda la soberbia en la que incurrimos cuando tratamos de fabricar criaturas inteligentes y sintientes a imagen y semejanza nuestras o de la mente divina. Rara vez sale bien. Desde el monstruo del *Frankenstein* de Mary Shelley hasta la Ava del filme *Ex_Machina* de Alex Garland, pasando por los robots epónimos de Karel Čapek, el *Terminator* de James Cameron, los replicantes de *Blade Runner* de Ridley Scott, o el HAL de Stanley Kubrick, estas creaciones casi siempre se vuelven contra sus creadores y dejan a su paso regueros de destrucción, melancolía y confusión filosófica.

Durante la última década, más o menos, el rápido auge de la IA ha

vuelto más urgente la necesidad dar respuesta a ciertas preguntas que nos hacemos acerca de la conciencia de las máquinas. Actualmente, estamos rodeados de IA: está integrada en nuestros teléfonos, en nuestros frigoríficos y en nuestros automóviles, y está impulsada en muchos casos por algoritmos de red neuronal inspirados en la arquitectura del cerebro. Hacemos bien en preocuparnos por el impacto de esta nueva tecnología. ¿Nos quitará nuestros puestos de trabajo? ¿Deshebrará el tejido de nuestras sociedades? ¿Y terminará por destruirnos a todos, ya sea por un incipiente interés propio, o por una falta de previsión programadora que lleve a que se agoten los recursos de la Tierra, transformados en una inmensa montaña de clips sujetapapeles?[2] Lo que subyace a muchas de esas preocupaciones, sobre todo a las de carácter más existencial y apocalíptico, es la suposición de que, en algún punto de su cada vez más acelerado desarrollo, la IA acabará por tener conciencia propia. Es el mito del golem hecho silicio.

¿Qué haría falta para que una máquina sea consciente? ¿Qué implicaciones tendría algo así? ¿Y cómo podríamos distinguir siquiera una máquina consciente de su equivalente zombi?

¿Por qué deberíamos *plantearnos* siquiera que una máquina –una inteligencia artificial– pueda llegar a tener consciencia? Como ya he mencionado, es un hecho muy habitual –aunque ni mucho menos universal– pensar que la conciencia surgirá de manera natural en las máquinas a partir del momento en que traspasen cierto umbral de inteligencia aún desconocido para nosotros. Pero ¿qué nos mueve a tener semejante intuición? Creo que dos son los supuestos clave responsables de la misma, y ninguno de los dos resulta justificable. El primero tiene que ver con las condiciones *necesarias* para que algo sea consciente. El segundo, con las condiciones *suficientes* para eso mismo.

El primer supuesto –el de la condición necesaria– es el *funcionalismo.* El funcionalismo dice que la conciencia no depende del material del que esté hecho un sistema, ya sea *wetware, hardware,* neuronas, puertas lógicas de silicio… o arcilla del río Moldava. El funcionalismo propugna

que lo que importa para la existencia de conciencia es lo que un sistema *hace*. Si un sistema transforma entradas *(inputs)* en salidas *(outputs)* de un modo correcto determinado, habrá conciencia. Como ya expliqué en el capítulo 1, ese postulado encierra dos tesis separadas. La primera es que el resultado es *independiente* del sustrato o material particular del soporte, mientras que la segunda es que las relaciones *input-output* son *suficientes* por sí solas para ese resultado. Para quienes suscriben postulados funcionalistas, esa independencia y esa suficiencia van unidas la mayoría de veces, pero también pueden aceptarse o rechazarse por separado.

El funcionalismo es una perspectiva popular entre los filósofos de la mente y tiende a ser aceptada como postura por defecto por muchos no filósofos también. Pero eso no quiere decir que sea correcta. Yo no veo argumentos irrebatibles ni a favor ni en contra de la postura de que la conciencia es independiente de su sustrato, o de que solo es una cuestión de relaciones entre *inputs* y *outputs,* es decir, de «procesamiento de información». Mi actitud ante el funcionalismo es de suspicaz agnosticismo.

Para que los ordenadores artificialmente inteligentes puedan ser conscientes, el funcionalismo tendría que ser verdad. Es la condición necesaria. Pero el hecho de que el funcionalismo fuese verdad no bastaría sin más: el procesamiento de información por sí solo no es condición suficiente de la conciencia. El segundo supuesto es que el tipo de información que por sí solo bastaría para la presencia de conciencia sea también el mismo que sirve de sustento a la inteligencia. Lo que se supone, a fin de cuentas, es que la conciencia y la inteligencia están estrechamente ligadas, incluso hasta el punto de ser constitutivas la una de la otra: la conciencia será una inevitable compañera de viaje de la inteligencia.

Pero no hay mucha evidencia en la que apoyar este otro supuesto. Como vimos en el capítulo anterior, la tendencia a fusionar la conciencia con la inteligencia se remonta en sus orígenes a un pernicioso

antropocentrismo que nos induce a sobreinterpretar el mundo a través de la distorsionadora óptica de nuestros propios valores y experiencias. *Somos* conscientes, *somos* inteligentes y nos sentimos tan orgullosos como especie de nuestra autoproclamada inteligencia que presuponemos que esta va inextricablemente unida a nuestra condición consciente y viceversa.

Pero, aunque la inteligencia permite una amplia ramificación de los estados de consciencia posibles para los organismos conscientes, es un error asumir que la inteligencia –al menos, en sus formas avanzadas– es necesaria o suficiente para la existencia de conciencia. Si insistimos en presuponer que la conciencia está intrínsecamente ligada a la inteligencia, podríamos dejarnos llevar en exceso por el ansia de atribuir conciencia a los sistemas artificiales que parecen inteligentes, y podríamos descartar demasiado pronto que la tengan otros sistemas (otros animales, por ejemplo) que no cuadren con nuestros cuestionables criterios humanos de competencia cognitiva.

En los últimos años, estos supuestos acerca de la necesidad y la suficiencia han quedado disimulados y desplazados por otra hornada de nuevas inquietudes y malentendidos que han vuelto más urgente y apocalíptica la (en el fondo) injustificada preocupación por la posibilidad de la llegada de una conciencia artificial.

He aquí unas cuantas de esas inquietudes. Existe la preocupación por que la IA –consciente o no– vaya ya camino de sobrepasar a la inteligencia humana y alejarse de nuestra capacidad de comprensión y de control. Es la llamada hipótesis sobre la «singularidad» tecnológica, popularizada por el futurólogo Ray Kurzweil y motivada por el extraordinario crecimiento de los recursos informáticos brutos a lo largo de las últimas décadas.[3] ¿En qué lugar de esa curva exponencial nos encontramos? El problema de las curvas exponenciales –como muchos aprendimos durante la reciente pandemia de coronavirus– es que sea cual sea el lugar en el que nos encontremos en ellas, lo que está por delante siempre nos parecerá increíblemente empinado, y lo que

hayamos dejado atrás nos resultará irrelevantemente plano. La visión local no nos da pistas de dónde estamos realmente. Tampoco podemos olvidar nuestro tradicional miedo prometeico a que nuestras creaciones se terminen volviendo contra nosotros de un modo u otro, un temor que un buen número de películas y libros de ciencia ficción han sabido detectar y reformular para infundírnoslo de nuevo con afán lucrativo. Y, por último, está el desgraciado hecho de que el término «conciencia» suele usarse y propagarse con perjudicial imprecisión cuando de las capacidades de las máquinas se trata. Para algunas personas –entre las que se incluyen algunos investigadores expertos en IA– cualquier cosa que responda a una estimulación, que aprenda algo o que se comporte de un modo que le permita maximizar una recompensa o alcanzar un objetivo es consciente. Para mí, eso no es más que estirar en exceso y sin sentido el significado razonable de la expresión «ser conscientes».

Difícilmente puede sorprendernos que todos esos ingredientes mezclados induzcan a muchas personas a pensar que la IA consciente está a la vuelta de la esquina y que debería preocuparnos mucho lo que ocurra cuando llegue. Cierto es que no es una posibilidad que se pueda descartar por completo. Si los propugnadores de la singularidad tecnológica terminan teniendo razón, esa preocupación estaría más que justificada. El problema es que, desde nuestra perspectiva actual, se trata de una posibilidad harto improbable. Muchas más probabilidades le podemos augurar, en cambio, a una situación como la ilustrada en la figura siguiente. Ahí vemos que la conciencia no viene determinada por la inteligencia, y que esta puede existir también sin conciencia.[4] Ambas se presentan en múltiples formas y se expresan en múltiples dimensiones diferentes, lo que significa que no hay una sola escala única ni para la conciencia ni para la inteligencia.[5]

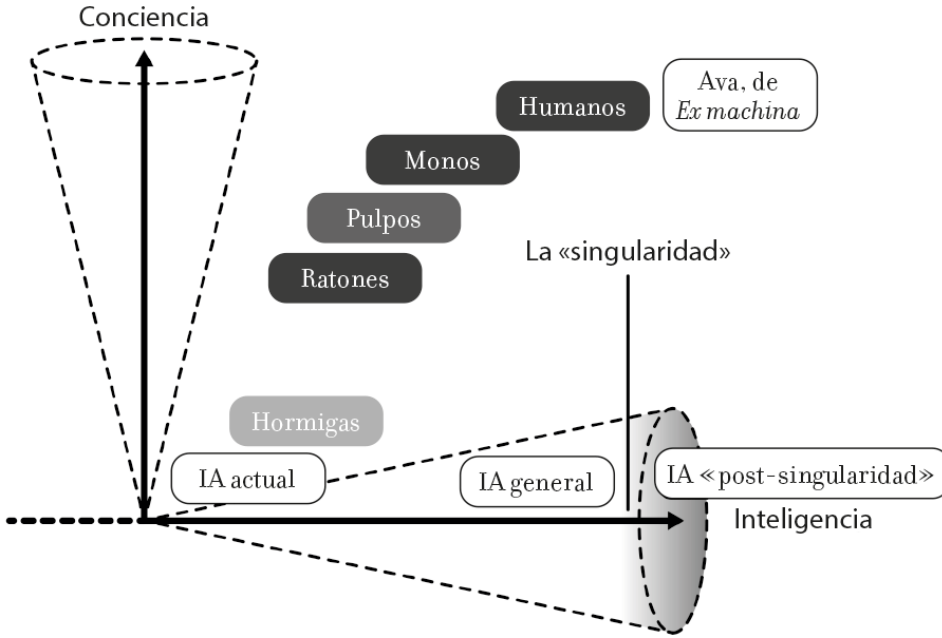

Fig. 21. La conciencia y la inteligencia son separables y multidimensionales. Las posiciones de todos estos animales y máquinas (reales e imaginarias) son meramente ilustrativas.

En esta representación, habrás notado que la IA actual está ubicada muy abajo en la escala de la inteligencia. Esto es así porque no está claro si los actuales sistemas de IA son inteligentes en un sentido significativo. Buena parte de la IA de hoy en día se podría describir como un reconocimiento sofisticado de patrones mediante el uso de máquinas, tal vez aderezado con un poco de planificación. E, inteligentes o no, estos sistemas hacen lo que hacen sin ser conscientes de nada.

En sus proyecciones de futuro, la gran ambición declarada de muchos investigadores expertos en IA es la de desarrollar sistemas con las capacidades de inteligencia general de un ser humano: lo que se conoce como la «inteligencia artificial general» o «IA general». Y más allá de ese punto, se extendería ya la *terra incognita* de la inteligencia «post-singularidad». Pero no hay un solo punto en ese recorrido en el que esté

justificado presuponer que la conciencia se sumará automáticamente al viaje. Es más, puede haber muchas formas de inteligencia que se desvíen de la de apariencia humana, y que complementen (en vez de sustituir o amplificar) el kit de herramientas cognitivas específico de nuestra especie, y todo eso, repito, sin que la conciencia esté implicada.

Puede que algunas formas concretas de inteligencia sean imposibles si no van acompañadas de conciencia, pero incluso si ese es el caso, no quiere decir que *todas* las formas de inteligencia –una vez superado cierto umbral todavía desconocido– requieran de una condición consciente concomitante. Sin embargo, sí es posible que todos los entes conscientes sean, como mínimo, un poco inteligentes, siempre que definamos la conciencia en un sentido lo bastante amplio. Pero, insisto, eso no significa que la inteligencia sea la ruta por excelencia hacia la conciencia.

No bastará con fabricar ordenadores más inteligentes para que algún día lleguen a ser sintientes. Pero eso no significa que la conciencia en máquinas sea imposible. ¿Qué pasaría si nos propusiéramos diseñarlas con conciencia desde el principio? Si no inteligencia, ¿qué se *necesitaría* para fabricar una máquina consciente?

La respuesta a esa pregunta depende de lo que consideres suficiente para que un sistema se pueda calificar como consciente, y eso dependerá a su vez de cuál sea la teoría de la conciencia que suscribas. No es de extrañar, por lo tanto, que haya tantas opiniones distintas sobre qué debería tener una máquina para ser consciente.

En el extremo más liberal del espectro se sitúan los que creen, muy en sintonía con el funcionalismo, que la conciencia solo depende de que se dé el tipo adecuado de procesamiento de información. No hace falta identificar ese procesamiento de información con la «inteligencia»; lo importante es que, tratándose de procesos y de contenidos informacionales, se puedan informatizar. De ahí que, por ejemplo, según una propuesta presentada en la revista *Science* en 2017, se pueda decir que una máquina es consciente si procesa información de un modo que

implique una «disponibilidad global» de dicha información y permita una «automonitorización» de su rendimiento. Los autores no aclararon si una máquina así *sería realmente* consciente o simplemente se *comportaría como si* lo fuera, pero la premisa subyacente es que no se necesita nada más para que haya conciencia que el hecho de que se produzca el tipo adecuado de procesamiento de la información.[6]

Mayor fuerza tiene aún la proposición referida a las máquinas conscientes que han formulado algunos defensores de la teoría de la información integrada (TII). Como vimos en el capítulo 3, desde la TII se afirma que la conciencia no es más que información integrada, y que la cantidad de información integrada producida por un sistema está totalmente determinada por propiedades de sus mecanismos internos, en concreto, por su «estructura de causas y efectos». Según la TII, cualquier máquina que genere información integrada, sea de lo que sea que esté hecha, y con independencia de cuál sea su aspecto externo, tendrá algún grado de conciencia. No obstante, la TII deja abierta también la posibilidad de que las máquinas parezcan conscientes a ojos de un observador externo, o inteligentes, o tal vez las dos cosas, pero que, al mismo tiempo, tengan mecanismos que no generen información integrada alguna y, por lo tanto, carezcan por completo de conciencia.[7]

Ninguna de esas teorías identifica la conciencia con la inteligencia, pero todas aceptan que, si unas máquinas satisficieran ciertas condiciones específicas –el tipo correcto de procesamiento de información, o la información integrada superior a cero– serían conscientes. Ahora bien, para aceptar estas implicaciones, antes, lógicamente, habría que aceptar las teorías en cuestión también.

La teoría del animal-máquina sobre la conciencia y el yo basa las experiencias que los seres conscientes tienen del mundo y del yo en el impulso biológico por conservar la integridad fisiológica (es decir, por seguir vivos). Pero ¿qué dice esta teoría acerca de la posibilidad de que haya máquinas conscientes?

Imagina un robot de un futuro cercano con un cerebro de silicio y un

cuerpo humanoide, equipado con toda clase de sensores y efectores. Este robot está controlado por una red neuronal artificial diseñada conforme a los principios del procesamiento predictivo y la inferencia activa. Las señales que fluyen por sus circuitos ponen en práctica un modelo generativo de su entorno y de su propio cuerpo. Continuamente está usando ese modelo para elaborar mejores conjeturas bayesianas acerca de las causas de sus *inputs* sensoriales. Estas alucinaciones controladas (y controladoras) de naturaleza sintética han sido diseñadas con la finalidad de mantener el robot en un estado funcional óptimo: para que siga «vivo» según su propio criterio. Incluso cuenta con *inputs* interoceptivos artificiales que le señalan sus niveles de batería y la integridad de sus actuadores y sus músculos sintéticos. Una serie de mejores conjeturas (orientadas al control) acerca de esos *inputs* interoceptivos generan unos estados emocionales sintéticos que motivan y guían la conducta del robot.

Este se comporta de manera autónoma y hace lo correcto en el momento oportuno para cumplir con sus objetivos. Actuando así, da la impresión a quienes lo observan desde fuera de que es un agente inteligente y sintiente. A nivel interno, sus mecanismos se corresponden de forma directa con el «mapa» de la maquinaria predictiva que, según he sugerido aquí, subyace a las experiencias humanas básicas de la corporeidad y la yoidad. Es un animal-máquina de silicona.

¿Sería consciente un robot así?

La respuesta honesta, aunque un tanto decepcionante, es que no lo sé con seguridad, pero probablemente no. La teoría del animal-máquina propugna que la conciencia en los seres humanos y en otros animales surgió con la evolución, que va apareciendo en cada uno de nosotros a lo largo de nuestro desarrollo vital, y que funciona a cada instante por vías estrechamente conectadas con nuestra condición de sistemas vivos. Todas nuestras experiencias y percepciones tienen su origen en nuestra naturaleza como máquinas vivas autosostenibles que se preocupan de su propia pervivencia. Yo intuyo (y, de nuevo, se trata solamente de una

intuición) que la materialidad de la vida terminará demostrándose muy importante para todas las manifestaciones de conciencia. Uno de los motivos por los que lo supongo es que el imperativo de la regulación y el automantenimiento en los sistemas vivos no se limita solamente a un nivel, como sería el de la integridad de todo el cuerpo. El automantenimiento para los seres vivos alcanza a sus niveles más bajos, incluso al de sus células individuales. Todas las células de tu cuerpo –de *cualquier* cuerpo– están regenerando constantemente las condiciones necesarias para su propia integridad en el tiempo.[8] No se puede decir lo mismo de ningún ordenador actual (o de un futuro próximo), y tampoco sería verdad para un animal-máquina de silicio como el que acabo de describir.

No cabe colegir de lo anterior que lo que estoy diciendo es que las células individuales son conscientes ni que todos los organismos vivos lo son. Lo único que quiero dejar claro es que los procesos de regulación fisiológica en los que se sustentan la conciencia y la yoidad en la teoría del animal-máquina reciben su impulso de ciertos procesos fundamentales de la vida que se manifiestan «en todos los niveles» de esta, desde el más alto hasta el más fundamental. Desde ese punto de vista, es la vida, más que el procesamiento de información, la que insufla un aliento especial a las fórmulas y las ecuaciones.

Pero incluso aunque las máquinas conscientes sean una posibilidad muy lejana –suponiendo que sean realmente posibles–, sigue habiendo mucho por lo que preocuparse. En un futuro más o menos inmediato, es perfectamente plausible que la evolución de la IA y de la robótica produzca nuevas tecnologías que nos den la impresión de ser conscientes, aun cuando no haya motivo concluyente alguno para creer que realmente lo *son*.

En la película *Ex_Machina,* dirigida por Alex Garland en 2014, Nathan, un genio de la tecnología multimillonario que lleva una vida muy recluida y solitaria, invita a Caleb, un as de la programación, a que venga a verlo en su remota guarida para que conozca a Ava, la robot

inteligente (y preguntona) que él ha creado. La misión de Caleb consiste en averiguar si Ava es consciente o bien no pasa de ser una robot inteligente sin vida interior alguna.

Ex_Machina está muy inspirada en el test de Turing, la famosa vara de medir si una máquina puede pensar o no.[9] En una escena especialmente incisiva, Nathan está examinando a Caleb con preguntas sobre esa prueba. En la versión estándar del test de Turing, como Caleb bien sabe, un juez humano interroga a distancia tanto a una máquina potencialmente inteligente como a otro humano, y lo hace intercambiando con ellos mensajes tecleados en un ordenador solamente. La máquina supera la prueba cuando el juez se muestra sistemáticamente incapaz de distinguir entre el humano y la máquina. Pero Nathan tiene algo mucho más interesante en mente. En el caso de Ava, dice, «la verdadera prueba es *mostrarte* que es un robot y ver si aún crees que tiene conciencia».

Este nuevo juego transforma el test de Turing, que pasa de ser un test de inteligencia a uno de conciencia, y, como ahora ya sabemos, la una y la otra son fenómenos muy diferentes. Además, Garland nos enseña que la prueba no tiene que ver en absoluto con la robot. Como dice Nathan, lo que importa no es si Ava es una máquina. Ni siquiera importa si Ava, aun siendo una máquina, tiene conciencia. De lo que se trata en realidad es de si Ava hace que una persona consciente crea que ella (el robot) es consciente también. Lo genial de esa conversación entre Nathan y Caleb es que revela lo que realmente son los test de ese tipo: una prueba para el ser humano, no para la máquina. Lo es tanto el test original de Turing como el equivalente propuesto por Garland en pleno siglo XXI para la detección de la conciencia. El diálogo de la película de Garland capta con tal elegancia la dificultad de atribuirle conciencia a una máquina que, en nuestro campo, el concepto «test de Garland» está cobrando una popularidad cada vez mayor: un raro ejemplo de retroalimentación de la ciencia ficción sobre la ciencia.[10]

Son ya muchos programas informáticos simples, incluidos *chatbots* de

varias clases, los que supuestamente han «superado» la prueba de Turing gracias a que una proporción suficiente de jueces humanos han sido engañados durante un tiempo juzgado suficiente también. Un ejemplo particularmente estrafalario de ese tipo de pruebas, también de 2014, fue uno en el que un *chatbot* que se hacía pasar por un niño ucraniano de trece años hizo creer a diez de un total de treinta jueces humanos que era un niño ucraniano de trece años de verdad. Aquello dio pie a que muchos proclamaran sin sonrojarse que por fin se habían superado ciertos hitos históricos en el terreno de la IA.[11] Pero, claro está, es más fácil imitar a un adolescente extranjero que habla mal el inglés que a alguien que sea de la misma edad, hable el mismo idioma y comparta la misma cultura que el juez humano de turno, sobre todo si solo están permitidas las interacciones remotas a través de mensajes de texto. Cuando el *chatbot* superó la prueba, su respuesta fue: «Me siento respecto a vencer el test de turing *[sic]* bastante apropiadamente».[12] Si bajamos el listón hasta ese punto, la prueba se vuelve mucho más fácil de superar. En realidad, lo que estaba a examen en un test como aquel era la credulidad humana y los humanos lo suspendieron.[13]

A medida que mejore la IA es posible que, muy pronto, el test de Turing se supere sin necesidad de reducir tan artificialmente los criterios. En mayo de 2020, el laboratorio de investigación OpenAI presentó GPT-3, una enorme red neuronal artificial entrenada con ejemplos de habla natural tomados de una amplísima muestra en internet.[14] Además de diálogos como los de los *chatbots,* GPT-3 puede generar pasajes sustanciales de texto en múltiples estilos diferentes cuando se le da pie con unas pocas palabras o frases iniciales. Aunque no comprende lo que produce,[15] la fluidez y la sofisticación del *output* de GPT-3 es sorprendente y algunos pueden encontrarla incluso inquietante. El diario *The Guardian* publicó un ejemplo de uno de los productos de dicha red: una redacción de quinientas palabras sobre por qué los seres humanos no deberían temer a la IA. En ella tocaba temas que iban desde la psicología de la violencia humana a la revolución

industrial, y decía cosas tan desconcertantes como que la IA no debería perder el tiempo tratando de entender los puntos de vista de personas que se ganan la vida desconfiando de la inteligencia artificial.[16]

Pese a su sofisticación, estoy bastante seguro de que un interlocutor humano mínimamente sofisticado todavía puede descubrir a GPT-3 con bastante facilidad. Tal vez ya no le sea posible cuando llegue GPT-4… o GPT-10. Pero incluso aunque un futuro sistema parecido a GPT lograse superar repetidamente la prueba de Turing con matrícula de honor, lo que estaría demostrando con ello no sería más que una forma muy limitada de inteligencia (simulada) –la relacionada con los intercambios lingüísticos incorpóreos–, muy alejada de la inteligencia natural plenamente corporeizada del «hacer lo que toca en el momento en que toca» que vemos en los seres humanos y en otros muchos animales (y que veríamos también en mi hipotético animal-máquina de silicio).

En lo relativo a la conciencia, no existe nada equivalente al *chatbot* ucraniano, ni menos aún a GPT (seguido del número que sea). El test de Garland continúa siendo un umbral virgen. De hecho, los intentos de crear simulacros de humanos sintientes han tendido a inducir en el público sensaciones de inquietud y repulsión,[17] muy alejadas de la compleja mezcla de atracción, empatía y lástima que Caleb sentía por Ava en *Ex_Machina*.

El robotista japonés Hiroshi Ishiguro lleva décadas fabricando robots lo más similares posible a seres humanos. Los llama «geminoides». Ishiguro ha creado copias geminoides de sí mismo (véase la página siguiente) y de su hija (cuando esta tenía seis años), así como una presentadora geminoide de televisión japonesa inspirada en una combinación de una treintena de personas distintas. Cada geminoide se fabrica a partir de copias escaneadas de cuerpos humanos en tres dimensiones y lleva integrados unos actuadores capaces de generar una amplia gama de expresiones faciales y gestos. Hay muy poca IA propiamente dicha en estos dispositivos robóticos: se basan en la imitación humana y una de sus posibles aplicaciones, entre otras, es la presencia remota (o

«telepresencia»). En una ocasión, Ishiguro usó su geminoide para impartir una clase a distancia de cuarenta y cinco minutos ante ciento cincuenta estudiantes universitarios de grado.

No se puede negar que los geminoides dan repelús. Son realistas, pero no lo bastante. Piensa en la diferencia entre encontrarte con un geminoide y encontrarte con un gato. Ante un gato –o un pulpo, da igual–, tenemos de inmediato la sensación de estar en presencia de otro ente consciente, por muy diferentes que sean nuestros aspectos visuales respectivos. Ante un geminoide, la impactante, a la vez que imperfecta, similitud física acentúa ciertas sensaciones de desconexión y de alteridad. Según un estudio de 2009, la sensación más común de aquellos que se encuentran frente a un geminoide es el miedo.

Este tipo de reacción es un ejemplo del llamado «valle de lo inquietante», un concepto inventado por otro investigador japonés, Masahiro Mori, en 1970. Según Mori, de entrada, los robots suscitan reacciones cada vez más positivas y empáticas en las personas a medida que los van construyendo más parecidos a ellas (pensemos en el C-3PO de *Star Wars,* por ejemplo). Sin embargo, hay un umbral en esa evolución a partir del cual pasan a parecer extraordinariamente humanos en ciertos sentidos, pero se quedan lejos de serlo en otros, y las reacciones de las personas ante ellos viran radicalmente hacia la repulsión y el miedo: la curva de la evolución de la aceptación humana de los robots entra así en el «valle de lo inquietante». Solo volverán a remontar hacia la aceptación positiva más adelante, a medida que el parecido logre ser tal que alcance el punto de lo indistinguible. Existen muchas teorías sobre por qué existe ese valle de lo inquietante, pero de lo que cabe poca duda es de que existe.[18]

Fig. 22. Hiroshi Ishiguro con su geminoide
Imagen utilizada con permiso de ATR Hiroshi Ishiguro Laboratories

Aunque cuesta mucho que los robots del mundo real superen esa depresión de lo inquietante, las novedades en el mundo virtual sí

parecen estar ya remontando la cara opuesta del valle y asomando por el otro lado. Algunos avances recientes en aprendizaje automático mediante el uso de «redes (neuronales) generativas antagónicas» (RGA) han podido generar rostros fotorrealistas de personas que nunca han existido (véase la página siguiente).[*] Estas imágenes se crean mezclando ingeniosamente rasgos tomados de grandes bases de datos de caras reales y empleando técnicas similares a las que usamos en nuestra máquina de alucinaciones (la que describimos en el capítulo 6). Cuando se combinan con tecnologías *deepfake,*[19] capaces de animar esos rostros para hacer que digan algo, y cuando lo que dicen viene producido por un *software* de reconocimiento de habla y de producción de lenguaje cada vez más sofisticado, como GPT-3, es como si de pronto nos encontráramos en un mundo poblado por personas virtuales que son indistinguibles en la práctica de las representaciones virtuales de personas reales. Es un mundo en el que terminaremos acostumbrándonos a no ser capaces de diferenciar quién es real de quién no lo es.

Quien piense que estos avances nunca llegarán a superar de forma convincente un test de Turing en formato vídeo porque chocarán antes con un tope insalvable andará probablemente bastante desencaminado. Ese modo de pensar es muy revelador de, o bien un caso resistente de excepcionalismo humano, o bien falta de imaginación, o bien de ambas cosas. Llegará el día en que superarán el test, seguro. Pero seguirán en pie dos interrogantes. El primero es si estas nuevas creaciones virtuales podrán traspasar la barrera del valle de lo inquietante con la que siguen chocando los geminoides de Ishiguro y entrar en el mundo real. El segundo es si también superarán el test de Garland. ¿Tendremos la sensación de que esas nuevas *agencias* son realmente conscientes, además de realmente inteligentes, aun cuando sepamos que no son más que líneas de comandos de un programa informático?

Y si nos llegamos a sentir así algún día, ¿cómo nos influirá eso?

Fig. 23. Ocho rostros. Estas personas no son reales.

El rápido auge de la IA –sea cual sea la mezcla de realidad y de bombo publicitario que lo esté alimentando– ha reavivado un debate ético renaciente y necesario. Muchas de las inquietudes que despierta la IA en el terreno ético tienen que ver con las consecuencias económicas y sociales de tecnologías que se prevén de aplicación bastante inminente, como los vehículos de conducción autónoma y los sistemas automatizados de producción fabril, que inevitablemente comportarán una importante alteración de estructuras y comportamientos.[*] Hay preocupaciones legítimas, también, relacionadas con la delegación de capacidades de toma de decisiones en sistemas artificiales cuyo funcionamiento interno puede estar sujeto a toda clase de sesgos y caprichos, y que podrían permanecer opacos, no ya a las personas afectadas, sino incluso a quienes los diseñaron. Llevado al extremo, ¿qué clase de horrores podrían llegar a desatarse si se pusiera un sistema de IA al frente de la gestión del armamento nuclear o de la columna vertebral de internet?

También generan preocupación ética las consecuencias psicológicas y conductuales de la IA y el aprendizaje automático. La invasión de la privacidad por medio de los *deepfakes,* la modificación de la conducta

por obra de los algoritmos predictivos, y la distorsión de creencias en las «burbujas de filtros» y las cámaras de resonancia que son las redes sociales, son solo unas pocas de las múltiples fuerzas que tensionan el tejido de nuestras sociedades y que nos empujan a ceder voluntariamente nuestras identidades y nuestra autonomía a unas grandes empresas de datos anónimas en lo que, en la práctica, constituye un inmenso experimento mundial incontrolado.[20]

Con semejante telón de fondo, los análisis éticos relacionados con la conciencia de las máquinas pueden parecer un tanto indulgentes y abstrusos. Pero no lo son. Son debates necesarios, aun si las máquinas en cuestión carecen (todavía) de conciencia. El día en que superen por fin el test de Garland, compartiremos nuestras vidas con entes que *sentiremos* que tienen su propia vida interior subjetiva, aunque sepamos (o creamos) que no la tienen. Cuesta prever las consecuencias psicológicas y conductuales de algo así. Es posible que aprendamos a distinguir entre cómo nos *sentimos* y cómo debemos *actuar* al respecto, hasta el punto de que nos parezca natural mostrar una preocupación y una consideración por el ser humano que no tendremos por el robot por mucho que sintamos que ambos tienen conciencia. De todos modos, no está claro qué consecuencias tendrá esto en nuestras psicologías individuales.

La serie de televisión *Westworld* nos muestra un futurista parque de atracciones donde se han desarrollado unos robots realistas con el fin específico de que los seres humanos puedan someterlos a toda clase de depravadas atrocidades, abusando de ellos, asesinándolos o violándolos. ¿Sería posible torturar a un robot sintiendo que es un ser consciente y sabiendo, al mismo tiempo, que no lo es, sin sufrir una fractura mental? Con las mentes que poseemos actualmente, un comportamiento así entraría directamente en la categoría de la sociopatía máxima. Pero también existe la posibilidad de que el círculo de nuestro interés y preocupación moral se vea distorsionado por nuestra tendencia antropocéntrica a experimentar mayor empatía hacia entes con los que

sentimos que tenemos una mayor similitud. Aunque, claro está, si esto fuera así, tal vez terminaríamos preocupándonos más por nuestros gemelos geminoides de nueva generación que por otros humanos, y no digamos ya que por otros animales.

Obviamente, no todos los futuros posibles son así de distópicos. Pero en plena carrera acelerada entre el progreso real de la IA y el bombo en torno a este, bueno será que vayamos haciendo hueco a una ética que esté bien fundamentada en los conocimientos de la psicología para que juegue un papel en todo esto. No basta con sacar una nueva tecnología al mercado y esperar a ver qué ocurre. Desde luego, no deberíamos perseguir a ciegas ese fin último que se le atribuye convencionalmente a la IA –el de reproducir la inteligencia humana hasta el punto de sobrepasarla. Tal como muy sensatamente escribió Daniel Dennett, estamos construyendo «herramientas inteligentes, no colegas», y debemos asegurarnos de reconocer la diferencia.[21]

Y luego está la posibilidad de que las máquinas tengan verdaderamente conciencia algún día. Si (voluntaria o involuntariamente) termináramos introduciendo en el mundo unas nuevas formas capaces de experiencia subjetiva, nos enfrentaríamos a una crisis ética y moral a una escala sin precedentes. Desde el momento en que algo tiene estatus consciente, también posee un estatus moral. Estaríamos obligados a minimizar su sufrimiento potencial del mismo modo que estamos obligados a minimizar el sufrimiento de las criaturas vivas (aunque no se puede decir que lo estemos haciendo demasiado bien en ese terreno). Además, en el caso de esos hipotéticos agentes artificialmente sintientes futuros, nos encontraríamos con la dificultad añadida de que tal vez no tuviésemos ni idea de qué tipos de conciencia estarían experimentando. Imagínate un sistema sujeto a una forma completamente nueva de sufrimiento para la que no se conoce equivalente en las personas, y que escapa a toda concepción y toda capacidad instintiva humana para comprenderla. Imagínate un sistema para el que ni siquiera fuera válida la distinción entre sentimientos

positivos y negativos, pues carece de la dimensión fenomenológica correspondiente con la que establecemos tal diferencia. El reto ético en ese caso radicaría en que ni siquiera sabríamos cuáles serían las cuestiones éticas relevantes.

Por muy lejana que siga estando la conciencia artificial real, su sola posibilidad remota ya merece nuestra atención. Aunque no sepamos aún qué se necesitaría para crear una máquina consciente, tampoco sabemos qué *no* se necesitaría.

En junio de 2019, el filósofo alemán Thomas Metzinger pidió la imposición inmediata de una moratoria de treinta años a todos los estudios dirigidos a generar lo que él llamó una «fenomenología sintética» precisamente por los motivos que acabo de exponer.[22] Yo estaba presente cuando hizo ese anuncio. Ambos éramos conferenciantes en un simposio sobre conciencia artificial organizado en Cambridge por el Centro Leverhulme para el Futuro de la Inteligencia. No es fácil seguir el contenido literal del ruego que Metzinger expuso aquel día. De hecho, según cómo se interprete, su alcance podría extenderse a buena parte (si no la totalidad) de la elaboración de modelos computacionales en el ámbito de la psicología. En cualquier caso, la esencia de su mensaje está clara: no deberíamos avanzar alegremente sin saber por dónde vamos, empeñados en crear conciencia artificial solo porque nos parezca interesante, útil o fascinante. No hay mejor ética que la ética preventiva.

Cuando el vitalismo estaba en su apogeo, seguramente habría parecido igual de ridículo hablar de la ética de la vida artificial como hoy nos lo pueda parecer hablar de una ética de la conciencia artificial. Pero aquí estamos, algo más de cien años después, con un conocimiento profundo de lo que hace posible la vida, además de con muchas herramientas novedosas con las que modificarla e incluso crearla. Disponemos de técnicas de edición génica como CRISPR, que dota a los científicos de capacidad para modificar fácilmente las secuencias de ADN y cambiar la función de los genes. Incluso contamos con la facultad de desarrollar organismos totalmente sintéticos, construidos desde los

propios genes «hacia arriba»: en 2019, unos investigadores de Cambridge crearon una variante de la *Escherichia coli* con un genoma cien por cien sintético. De pronto, la ética de la creación de nuevas formas de vida adquirió plena relevancia.[23]

Y quién sabe si no será la biotecnología, más que la IA, la que más cerca nos lleve de la conciencia sintética. En este terreno, es particularmente significativa la llegada de los «organoides cerebrales». Se trata de minúsculas estructuras parecidas a cerebros, integradas por neuronas reales, que se cultivan a partir de células madre humanas pluripotentes (es decir, células que pueden diferenciarse en múltiples formas distintas). No son «minicerebros», pero los organoides cerebrales guardan semejanzas con el cerebro humano en desarrollo que los pueden hacer útiles como modelos de laboratorio de patologías que obedecen a un mal desarrollo cerebral. ¿Podrían esos organoides albergar alguna forma primitiva de consciencia incorpórea? Es difícil descartar tal posibilidad, sobre todo cuando se sabe que han empezado a mostrar ondas coordinadas de actividad eléctrica no muy diferentes de las que se aprecian en los bebés humanos prematuros, según un estudio reciente.[24]

A diferencia de los ordenadores, los organoides cerebrales están hechos del mismo material físico que los cerebros reales, lo que elimina un posible obstáculo para imaginar que puedan ser conscientes. Por otra parte, son todavía sumamente simples, son del todo incorpóreos y no interactúan para nada con el mundo exterior (aunque se los puede interconectar con cámaras, brazos robóticos y dispositivos parecidos). Pero, aun así, y pese a que la probabilidad de que los organoides actuales lleguen a ser conscientes algún día es muy remota, la pregunta estará cada vez más desconcertantemente abierta cuanto más se desarrolle esa tecnología.[25] Y eso nos lleva de vuelta a la necesidad de una ética preventiva. La posibilidad de conciencia en los organoides nos apremia desde el punto de vista ético no solo porque no se puede descartar, sino también por la escala potencial a la que se podría

materializar.[26] De hecho, ya lo ha anunciado el investigador especializado en organoides Alysson Muotri: «Queremos crear granjas de estos organoides».[27]

¿Por qué nos atrae tanto la posibilidad de que las máquinas tengan conciencia? ¿Por qué ejerce semejante tirón gravitacional sobre nuestra imaginación colectiva? Al final, pienso que tiene que ver con una especie de arrebato tecnológico nuestro, un deseo muy arraigado de trascender nuestra existencia biológica, restringida y desaliñadamente material, a medida que se acerca nuestro momento final. Si las máquinas conscientes son posibles, nace con ellas la posibilidad también de realojar nuestras mentes conscientes desde su actual soporte de *wetware* hacia la prístina circuitería de un futuro superordenador que nunca envejezca ni muera. Entraríamos así en el territorio de la carga y la descarga de mentes, uno de los temas favoritos de futurólogos y transhumanistas, para quienes una vida no es suficiente.[28]

Hay quienes piensan incluso que tal vez estemos ya allí. En su «hipótesis de la simulación», Nick Bostrom, profesor de Filosofía de la Universidad de Oxford, expone un argumento estadístico, según el cual, es más probable que formemos parte de una simulación informática enormemente sofisticada, diseñada e implementada por unos descendientes nuestros (muy superiores a nosotros a nivel tecnológico y, al parecer, obsesionados por la genealogía), que no que sigamos perteneciendo a la raza humana biológica original. Según esa hipótesis, ahora mismo ya seríamos agentes sintientes virtuales en un universo virtual.[29]

Algunos de los cautivados por el arrebato tecnológico ven la llegada inminente de una «singularidad», un punto crítico en la historia en el que la IA estará lista para independizarse más allá de nuestro control y de nuestra comprensión. En el mundo posterior a la «singularidad», cabe suponer que abundarán las máquinas conscientes y las simulaciones de antepasados. Nosotros, formas de vida basadas en carbono, nos quedaremos muy rezagados; nuestro momento de gloria se habrá ido

para no volver.

No hace falta una especial perspicacia sociológica para apreciar lo atractivo que este embriagador brebaje debe de ser para nuestra élite tecnológica, autoproyectada así en el papel de actor capital de esa transición sin precedentes en la historia humana, con la inmortalidad como premio final. Es lo que ocurre cuando el excepcionalismo humano se desboca. Visto desde esta perspectiva, todo ese alboroto en torno a la conciencia de las máquinas es sintomático de nuestra alienación creciente con respecto a nuestra naturaleza biológica y nuestra herencia evolutiva.

La perspectiva del animal-máquina difiere de este otro relato en casi todos los sentidos. Desde el punto de vista de mi teoría, como ya hemos visto, la totalidad de la experiencia y la vida mental humanas surge debido a (y no a pesar de) nuestra naturaleza como organismos biológicos autosustentados que se preocupan de su propia pervivencia. Esta visión de la conciencia y de la naturaleza humana no excluye la posibilidad de que haya máquinas conscientes, pero sí socava la validez de ese relato –henchido de arrebato tecnológico– que nos habla de la llegada inminente de ordenadores sintientes y suscita nuestros temores al tiempo que impregna nuestros sueños. Desde la perspectiva del animal-máquina, el propósito de llegar a comprender mejor la conciencia no nos aleja de la naturaleza, sino que nos va situando cada vez más dentro de su terreno.

Como debe ser.

EPÍLOGO

> Quiero tener control,
> quiero un cuerpo perfecto,
> quiero un alma perfecta.

RADIOHEAD, «Creep» (1992)

En enero de 2019, tuve directamente ante mis ojos un cerebro humano vivo por primera vez en mi vida. Fue veintipocos años después de que empezara a investigar la ciencia de la conciencia, diez años después de la inauguración de nuestro laboratorio en Sussex, y tres años después de aquel olvido y aquella inconsciencia inducidos en mí por la anestesia con los que comencé este libro. Tras todo ese tiempo, al contemplar aquella superficie cortical grisáceo-blanquecina tenuemente latiente, delicadamente hilada de venas rojas oscuras, volvió a parecerme inconcebible que un montón de materia así pueda dar origen a todo un mundo interior de pensamientos, sentimientos y percepciones: una vida vivida enteramente en primera persona. La profunda sensación de asombro que sentí se mezcló incómodamente en mi mente con el viejo chiste de la profesión, según el cual un trasplante de cerebro es la única operación en la que es mejor ser el donante que el receptor.

Mi anfitrión de ese día fue Michael Carter, un cirujano cerebral pediátrico que trabaja en el hospital infantil Bristol Royal, en el oeste de Inglaterra. Me había invitado para que observara uno de los procedimientos quirúrgicos más espectaculares que se realizan actualmente en el mundo. El paciente, un niño de apenas seis años, tenía programada una hemisferectomía. Sufría una epilepsia grave de nacimiento. El origen de sus ataques estaba en el hemisferio cortical

derecho, que había quedado seriamente dañado durante su parto, que fue prematuro. Todas las medicaciones convencionales con las que se había intentado pararlos habían sido en vano, así que hubo que acudir al neurocirujano como último recurso.

La hemisferectomía consiste en la desconexión neural completa de ese hemisferio cerebral derecho disfuncional. El cirujano se introduce en el cerebro por ese lado derecho, retira («extirpa») el lóbulo temporal y, a continuación, secciona todos los haces de conexiones –los tractos de materia blanca– que enlazan el hemisferio derecho con el resto del cerebro y del organismo. El hemisferio así aislado se queda dentro del cráneo y continúa conectado con el riego sanguíneo. Es una isla viva –aunque aislada– de corteza cerebral.[1] Esta versión extrema de una operación de cerebro dividido se basa en la idea de que la desconexión neural completa impide que las tormentas eléctricas que se originan en el lesionado hemisferio derecho se extiendan al resto del encéfalo. Si la operación se realiza a una edad lo bastante temprana, el cerebro joven suele mostrar una adaptabilidad suficiente como para que el hemisferio restante compense la mayor parte (o la totalidad) de la capacidad inicialmente perdida. Pese a la radicalidad de esta cirugía, y aunque cada caso es diferente, los resultados por lo general son buenos.

Esta operación en concreto comenzó hacia el mediodía y duró algo más de ocho horas. En mi trabajo, apenas puedo estar cinco minutos sin que me distraiga un correo electrónico, o el marcador de un partido de críquet, o las ganas de prepararme otra taza de té. Michael trabajó sin pausa, hora tras hora, de forma paciente, metódica, constante, asistido por un neurocirujano en prácticas y un equipo rotatorio de ayudantes. Hacia la mitad del procedimiento, en un momento en que el cirujano asistente se tomó un breve descanso, me invitó a que ocupara su sitio y me acercara al microscopio quirúrgico. No me esperaba disponer de tan privilegiada vista. Mientras miraba el interior de las iluminadas cavidades del cerebro del pequeño, traté de trasladar mi conocimiento abstracto de las diferentes regiones y vías al amasijo de alumbrado

tejido que tenía ante mí. Pero poco sentido tenía aquel ejercicio. No se apreciaba rastro visible alguno de las nítidas jerarquías corticales ni del hilado a contraflujo de señales ascendentes y descendentes que conocía por mis investigaciones. El cerebro volvía a ser de nuevo inescrutable para mí, y me quedé admirado tanto de la habilidad del neurocirujano como de la realidad material de aquel objeto más mágico que ninguno. Me sentía casi como un transgresor. Alguien había corrido un telón y había destapado algo demasiado íntimo como para estar tan a la vista. Y allí estaba yo, observando directamente la mecánica del yo humano.
La cirugía salió según lo previsto. Poco después de las ocho, Michael dejó que el cirujano en prácticas se encargara de terminar de recoser el cuero cabelludo del paciente y me llevó con él para que conociera a los familiares del niño. Estaban agradecidos y aliviados. Yo me preguntaba cómo se sentirían si hubieran visto lo que yo había visto aquel día.

Más tarde, mientras conducía de vuelta a casa abriéndome paso entre la temprana oscuridad exterior invernal, mis pensamientos volvieron a la definición del problema difícil de la conciencia según David Chalmers: «Existe un amplio consenso en torno al hecho de que la experiencia nace de una base física, pero no tenemos una buena explicación de por qué ni de cómo nace. ¿Por qué el procesamiento físico da origen a una rica vida interior? No parece objetivamente razonable que lo haga, y, aun así, lo hace».[2]

La filosofía ha dado respuestas optativas a este misterio que han ido desde el pampsiquismo (la idea de que hay conciencia en todas partes, más o menos) hasta el materialismo eliminatorio (no hay conciencia o, al menos, no como nos la imaginamos), pasando por toda clase de propuestas intermedias. Pero la ciencia de la conciencia no consiste en elegir una opción entre un menú de ellas, por muy de postín que sea el restaurante o por muy diestro que sea el chef en lo suyo. Se parece más bien a cocinar con lo que tengas en el frigorífico, y combinar y recombinar de diferentes modos y formas trozos y pedazos diversos de filosofía, neurociencia, psicología, informática, psiquiatría, aprendizaje

automático, etcétera, hasta convertirlos en algo nuevo.

Esa es la esencia del estudio de la conciencia desde el enfoque del problema real. Primero, aceptamos que la conciencia existe y, luego, nos preguntamos cómo se relacionan las diferentes propiedades fenomenológicas de esa conciencia –es decir, la manera en que están estructuradas las experiencias conscientes, la forma que adoptan– con las propiedades del cerebro, un cerebro que está corporeizado, inserto tanto en un organismo como en un mundo. Podemos empezar a responder a esas preguntas detectando correlaciones entre patrones concretos de actividad cerebral y tipos específicos de experiencia consciente, pero nuestras respuestas no tienen por qué limitarse a eso y, de hecho, no deberían quedarse solo ahí. El reto consiste en tender puentes explicativos cada vez más robustos entre los mecanismos y la fenomenología, de tal modo que las relaciones que establezcamos sean lo menos arbitrarias posible y tengan el máximo sentido. ¿Y qué significa «tener sentido» en este contexto? Ya lo hemos dicho aquí: explicar, predecir y controlar.

Esta estrategia se hace eco histórico del modo en que nuestro conocimiento científico de la vida trascendió el pensamiento mágico del vitalismo individuando las propiedades de los sistemas vivos para, a continuación, explicar cada una de ellas en términos de sus mecanismos subyacentes. La vida y la conciencia son cosas distintas, desde luego, pero a estas alturas espero haber convencido al lector de que están más estrechamente ligadas la una a la otra de lo que podría parecer en un principio. Sea como sea, la estrategia es la misma. En vez de tratar directamente de solucionar el problema difícil de la conciencia, y en lugar de dejar completamente de lado sus cualidades experienciales (las de la conciencia, quiero decir), el enfoque del problema real ofrece una esperanza auténtica de reconciliar lo físico con lo fenoménico, disolviendo así el problema difícil en vez de resolverlo.

Nuestro punto de partida fue el *nivel* consciente –la diferencia entre estar en coma y estar plenamente despierto y lúcido– y nos centramos

entonces en la importancia de la *medición*. La clave ahí es que hay unos buenos indicadores potenciales –como la densidad causal y la información integrada– que lo son por razones en absoluto arbitrarias. Logran captar unas propiedades muy presentes en *todas* las experiencias conscientes, como, por ejemplo, el hecho de que toda experiencia consciente sea unificada y, a la vez, distinta de todas las demás. Toda escena consciente se experimenta «de una pieza», y toda experiencia es como es y no de ningún otro modo.

Luego pasamos a la naturaleza del *contenido* consciente y, en particular, a la experiencia de ser un *yo* consciente. Planteé entonces una serie de desafíos a cómo nos parece que son las cosas; desafíos que, uno tras otro, nos iban instando a adoptar nuevas perspectivas «poscopernicanas» sobre la percepción consciente.

El primero de esos desafíos fue entender la percepción como una construcción activa, orientada a la acción, más que como el registro pasivo de una realidad externa objetiva. Nuestros mundos percibidos son menos y, al mismo tiempo, más que esa realidad externa objetiva (sea lo que sea esta). Nuestros cerebros crean nuestros mundos por medio de procesos de formación de mejores conjeturas bayesianas en los que las señales sensoriales sirven principalmente para refrenar o controlar nuestras hipótesis perceptivas, que no dejan de evolucionar. Vivimos dentro de una alucinación controlada que la evolución ha diseñado con el fin no de la precisión, sino de la utilidad.

El segundo desafío proyectó esa misma idea hacia dentro, hacia la experiencia de ser un yo. Exploramos cómo el yo es él mismo una percepción, una variedad más de alucinación controlada. Desde las experiencias de la identidad personal y de la continuidad en el tiempo, hasta el vago sentimiento de ser un cuerpo vivo, sin más, estos pedazos-de-yoidad dependen todos de la misma danza delicada entre la predicción perceptiva que procede de dentro hacia fuera y el error de predicción que procede de fuera hacia dentro, solo que, en este caso, buena parte de esa coreografía se desarrolla dentro de los confines del

propio cuerpo.

El desafío final consistía en comprender que la maquinaria predictiva de la percepción consciente tiene su razón de ser original y su función primaria no en la representación del mundo o del cuerpo, sino en el control y la regulación de nuestra condición fisiológica. Lo que moldea la totalidad de nuestras percepciones y cogniciones –el panorama integral de la experiencia humana y la vida mental– es un impulso biológico arraigado que nos mueve a seguir vivos. Percibimos el mundo que nos rodea, y nos percibimos a nosotros mismos en él, *con, a través de y debido a* nuestros cuerpos vivos.

Esa es mi teoría del animal-máquina, una versión (o inversión) de *l'homme machine* de Julien Offray de La Mettrie llevado al siglo XXI. Y ahí es donde se introducen los cambios más profundos en la concepción de la conciencia y la yoidad.

Siempre ha sido un misterio, por ejemplo, que las experiencias de «ser un yo» resulten tan diferentes de las experiencias que tenemos del mundo que nos rodea. Con esta teoría, podemos entender unas y otras como distintas expresiones de los mismos principios de la predicción perceptiva; sus diferencias en fenomenología tienen su origen en los diferentes tipos de predicción implicados: unas inferencias perceptivas se orientan a hacer averiguaciones sobre objetos presentes en el mundo exterior, mientras que las otras tienen que ver con el control del interior del organismo.

Ligando nuestras vidas mentales a nuestra realidad fisiológica, damos una sustancia renovada a ciertas concepciones tradicionales sobre la continuidad entre la vida y la mente, apuntaladas con los robustos pilares del procesamiento predictivo y del principio de la energía libre. Y esta continuidad profunda permite que, al mismo tiempo, nos concibamos en una relación más estrecha con los demás animales y con el resto de la naturaleza, y a la vez, más alejada de los cálculos incorpóreos de la IA. De este modo, por un lado, se vinculan la conciencia y la vida, mientras que, por el otro, se separa la conciencia

de la inteligencia. Esta reorientación de nuestro lugar en la naturaleza es aplicable no solo a nuestros cuerpos físicos, biológicos, sino también a nuestras mentes conscientes, a nuestras experiencias del mundo que nos rodea y del ser quienes somos.

Cada vez que la ciencia nos ha desplazado del centro de las cosas, nos ha dado mucho más a cambio. La revolución copernicana nos brindó un universo: uno en el que los descubrimientos astronómicos de los últimos cien años han superado con mucho los límites de la imaginación humana. La teoría de Charles Darwin de la evolución por medio de la selección natural nos dio una familia, un vínculo con todas las demás especies vivas y una mejor apreciación del tiempo profundo y del poder del diseño evolutivo. Y ahora la ciencia de la conciencia, de la que la teoría del animal-máquina no es más que una parte, está traspasando la última barrera defensiva que le quedaba al excepcionalismo humano (es decir, a la presuposición de que nuestras mentes conscientes son únicas y especiales) y nos está mostrando que también esto está profundamente integrado en los patrones generales de la naturaleza.

Todo en la experiencia consciente es percepción de algún tipo, y toda percepción es una alucinación controlada (o controladora) de alguna clase. Lo que más me entusiasma de este modo de pensar es lo lejos que puede llevarnos. Las experiencias individuales de libre albedrío son percepciones. El flujo del tiempo es una percepción. Puede que incluso la estructura tridimensional del mundo tal como lo experimentamos sea un aspecto perceptual, y que también lo sea la sensación de que los contenidos de la experiencia perceptiva son objetivamente reales.[3] Las herramientas de la ciencia de la conciencia nos están permitiendo acercarnos cada vez más al noúmeno de Kant, esa realidad incognoscible en última instancia de la que también nosotros formamos parte. Todas estas ideas son contrastables y, sea hacia donde sea que apunten los resultados de las pruebas, el simple hecho de que nos planteemos preguntas de este cariz reconfigura por sí solo nuestra concepción de lo que es la conciencia, de cómo se produce y de para qué sirve. Cada paso

que damos en esa dirección va minando un poco más la tan seductora como escasamente útil idea de que la conciencia es un gran misterio siniestro en busca de una solución igual de grande y aterradora.

Son muchas, también, las implicaciones prácticas de todo esto. Los indicadores que se utilizan para medir el nivel consciente, inspirados en toda esa base teórica, están contribuyendo a la implantación de nuevos «medidores» de la conciencia que cada vez son más capaces de detectar hasta la consciencia residual –o «encubierta»– que permanece en algunos pacientes que no manifiestan ninguna respuesta conductual. Los modelos informáticos de la percepción predictiva están arrojando luz sobre la base de las alucinaciones y los delirios, y están inaugurando así una transformación en el campo de la psiquiatría, que deja de centrarse tanto en el tratamiento de los síntomas para abordar también las causas. Y son múltiples y muy variadas las nuevas direcciones que están tomando la IA, las interfaces cerebro-máquina y la realidad virtual, entre otras muchas tecnologías, tanto consolidadas como emergentes. Indagar en la base biológica de la conciencia está resultando sorprendentemente útil.

Dicho esto, no está de más reconocer que enfrentarse al misterio de la consciencia es –y siempre será– un viaje muy personal. ¿De qué sirve una ciencia de la conciencia si no es para arrojar luz sobre nuestras vidas mentales individuales y sobre las vidas interiores de quienes nos rodean?

Ese es el aspecto verdaderamente prometedor del enfoque del problema real. Nos lleve adonde nos lleve finalmente, seguir ese camino nos ayudará a comprender un sinfín de cosas nuevas acerca de las experiencias conscientes del mundo que nos rodea y de nosotros mismos dentro de ese mundo. Veremos cómo nuestro universo interior es una parte (y no un aparte) de la naturaleza. Y aunque tal vez no pensemos en ello tan a menudo como podríamos, tendremos la oportunidad de estar en paz de un modo nuevo y diferente con lo que ocurre –o, mejor dicho, deja de ocurrir– cuando llega ese momento final de nuestra vida

en que la alucinación controlada de *ser tú* se disuelve en la nada: cuando el olvido no es una interrupción (inducida por la anestesia) del río de la conciencia, sino un retorno a la eternidad de la que cada uno de nosotros emergió en algún momento.

Aunque, bien pensado, tal vez no esté tan mal que, al final de *esta* historia, cuando la vida en primera persona alcanza su momento concluyente, siga habiendo un poco de misterio.

AGRADECIMIENTOS

Las ideas recogidas en este libro se han ido formando en las incontables conversaciones que he mantenido con amigos, colegas, alumnos, profesores y mentores durante más de dos décadas. Gracias a todos.

Gracias también a todos los miembros pasados y presentes de mi grupo de investigación en Sussex. Ha sido (y continúa siendo) un privilegio trabajar con vosotros. Y, en concreto, estoy particularmente agradecido a aquellos miembros en cuyas investigaciones me he basado para escribir este libro. Gracias, Adam Barrett, Lionel Barnett, Peter Lush, Alberto Mariola, Yair Pinto, Warrick Roseboom, Michael Schartner, David Schwartzman, Maxine Sherman, Keisuke Suzuki y Alexander Tschantz.

Asimismo, gracias, Manuel Baltieri, Reny Baykova, Luc Berthouze, Daniel Bor, Chris Buckley, Acer Chang, Paul Chorley, Ron Chrisley, Andy Clark, Marianne Cole, Clémence Compain, Guillaume Corlouer, Hugo Critchley, Zoltán Dienes, Tom Froese, Paul Graham, Inman Harvey, Owen Holland, Ryota Kanai, Tomasz Korbak, Isabel Maranhão, Federico Micheli, Beren Millidge, Thomas Nowotny, Andy Philippides, Charlotte Rae, Colin Reveley, Ryan Scott, Lina Skora, Nadine Spychala, Marta Suárez-Pinilla, Chris Thornton, Hao-Ting Wang y Jamie Ward, colegas y amigos (todos y todas) de Sussex en el pasado y en el presente. Vuestro trabajo y vuestras ideas han sido una gran contribución a las mías propias, y muchos habéis tenido la gran amabilidad de leer y comentar partes de este libro.

Estoy hondamente agradecido de haber tenido como mentores a científicos excepcionales en todas las etapas de mi trayectoria. Nicholas Mackintosh, que falleció en 2015, me supervisó durante mis estudios de

grado en Cambridge y me transmitió la confianza para iniciarme en la profesión académica. Phil Husbands dirigió mi doctorado y mi tesis en Sussex y me concedió libertad para explorar al tiempo que vigilaba de cerca hacia dónde me encaminaban mis pasos. Gerald Edelman, fallecido en 2014, también fue mentor mío durante mis más de seis años como investigador posdoctoral. Bajo su guía, mi interés por la conciencia terminaría convirtiéndose en mi trabajo. También me gustaría dar las gracias a Margaret Boden, Andy Clark, Daniel Dennett y Thomas Metzinger, que han sido todos extraordinarias fuentes de influencia y de inspiración para mí durante muchos años.

Varias personas asumieron la labor de leer y comentar en detalle los primeros borradores completos de este libro. Quiero expresar mi más sincero agradecimiento a Tim Bayne, Andy Clark (otra vez), Claudia Fischer, Jakob Hohwy y Murray Shanahan. Me he beneficiado inmensamente de sus ideas y sugerencias. También les estoy profundamente agradecido a Karl Friston, Marcello Massimini, Thomas Metzinger (otra vez), Aniruddh (Ani) Patel y Adrian Owen por sus detallados consejos sobre diversas partes concretas. Muchísimas gracias a Steve West (Lazy Chief) por sus elegantes ilustraciones. Vaya mi agradecimiento, también, para Baba Brinkman, cuyo talento con las palabras es una inspiración continua para mí. Trabajar con él en su *Rap Guide to Consciousness* fue un verdadero punto álgido en mi carrera. Gracias a Michael Carter por llevarme a su quirófano para ser testigo directo de una hemisferectomía. Y a David Edelman y Graziano Fiorito por introducirme en el mundo de los pulpos, y a Ian Cheney y Laurie Santos por hacer lo propio con el de los monos, y a Luise Schreiter por conducirme a otro mundo completamente diferente.

De las muchas otras personas cuyos consejos, ideas y apoyo me han ayudado a lo largo de muchos años, solo puedo mencionar aquí a unas pocas. Gracias, Anil Ananthaswamy, Chris Anderson, Bernard Baars, Lisa Feldman Barrett, Isabel Behncke, Tristan Bekinschtein, Yoshua Bengio, Heather Berlin, Matt Bergman, David Biello, Robin Carhart-Harris,

Olivia Carter, David Chalmers, Craig Chapman, Axel Cleeremans, Athena Demertzi, Steve Fleming, Zaferios Fountas, Friday Football, Chris Frith, Uta Frith, Alex Garland, Mariana Garza, Mel Goodale, Annaka Harris, Sam Harris, Nick Humphrey, Rob Illife, Robin Ince, John Iversen, Eugene Izhikevich, Alexis Johansen, Robert Kentridge, Christof Koch, Sid Kouider, Jeff Krichmar, Victor Lamme, Hakwan Lau, Steven Laureys, Rafael Malach, Daanish Masood, Simon McGregor, Pedro Mediano, Lucia Melloni, Liad Mudrick, Phil Newman, Angus Nisbet, the Parasites, Megan Peters, Giovanni Pezzulo, Tony Prescott, Blake Richards, Fernando Rosas, Adam Rutherford, Tim Satterthwaite, Tom Smith, Narayanan Srinivasan, Catherine Tallon-Baudry, Giulio Tononi, Nao Tsuchiya, Nick Turk-Browne, Lucina Uddin, Simon van Gaal, Bruno van Swinderen, Anniek Verholt, Paul Verschure, Lucy Walker, Nigel Warburton, Lisa Westbury y Martijn Wokke.

Estas personas han enderezado mi rumbo en múltiples sentidos a lo largo de los años. Eso no quiere decir que estén de acuerdo con todo lo que digo. Todos los errores o idioteces que continúen figurando en el libro son, por supuesto, responsabilidad mía al cien por cien.

No podría haber escrito esta obra sin el apoyo de la Fundación Wellcome (gracias a una beca Engagement), la Fundación Dr. Mortimer y Theresa Sackler, y el Instituto Canadiense de Investigaciones Avanzadas (de cuyo programa sobre Cerebro, Mente y Conciencia soy codirector). Les estoy agradecido por haber apoyado tanto mi tiempo como la investigación que realiza mi grupo. Y gracias a la Universidad de Sussex por procurarme un hogar académico desde hace ya muchos años.

Mi agradecimiento final lo reservo para mi agente y mis editores. Will Francis, de Janklow & Nesbit, me animó a escribir este libro, para empezar, y luego me ayudó a dar forma a la propuesta, a hallar excelentes editores, y a guiar en general la totalidad del proyecto desde el principio hasta el final. Gracias, igualmente, a esos excelentes editores de los que hablaba, Rowan Cope y Laura Hassan (de Faber & Faber), a

Stephen Morro (de Dutton/Penguin) y a Eleanor Rees como revisora.

NOTAS

PREFACIO

1.

Barnes (2008).

1. EL PROBLEMA REAL

1.

Baars (1988); Dehaene y Changeux (2011); Mashour *et al.* (2020); Shanahan (2010).

2.

Otra manera de entender la teoría del espacio de trabajo global es concibiéndola como una teoría de la «conciencia de acceso», por contraste con la «conciencia fenoménica». La conciencia fenoménica alude explícitamente a la experiencia. La conciencia de acceso enfatiza más bien las funciones cognitivas. Cuando un estado mental es «de acceso consciente», significa que está disponible para toda clase de funciones cognitivas, incluidas las de razonamiento, toma de decisiones y control de la conducta. Véase Block (2005).

3.

Hay múltiples variedades de teorías del pensamiento de orden superior. Véanse R. Brown *et al.* (2019); Fleming (2020); Lau y Rosenthal (2011).

4.

Portin (2009).

5.

Chalmers (1995a), p. 201.

6.

En fecha más reciente, Chalmers ha introducido también el «metaproblema» de la conciencia, que es el problema de por qué las personas piensan que existe un problema difícil de la conciencia (Chalmers, 2018). El metaproblema es, en realidad, una versión del problema fácil, ya que consiste en explicar una conducta (en este

caso, la conducta verbal de las personas que expresan su creencia de que existe un problema difícil de la conciencia). Una de las cosas que me gustan del metaproblema es que se puede reconocer que es un problema y estudiarlo con independencia de cuál sea la postura que se haya adoptado acerca de la conciencia en sí.

7.

Craver y Tabery (2017).

8.

Chalmers (1995a), p. 203 (la cursiva es la del original).

9.

La diferencia entre fisicalismo y materialismo es fundamentalmente histórica. El materialismo es una manera más antigua de llamar al fisicalismo. Hay quienes sostienen también que el fisicalismo es una «hipótesis lingüística» (la de que todo enunciado lingüístico equivale a un enunciado físico), mientras que el materialismo es una tesis más general sobre la naturaleza de las cosas. Véase Stoljar (2017).

10.

Aunque la mayoría de los funcionalistas son fisicalistas, también es posible ser funcionalista sin ser fisicalista.

11.

En su absorbente *The Idea of the Brain* (2020), Matthew Cobb traza el recorrido histórico de cómo la funcionalidad del cerebro se ha interpretado en cada época según la tecnología dominante en aquel momento (y, a veces, a la inversa).

12.

La historia del programa original, AlphaGo, está maravillosamente bien contada en una película homónima, en <**https://www.alphagomovie.com**>. Habrá quien bromee diciendo que estos programas, más que al go en sí, juegan a «la historia del go».

13.

John Searle ha elaborado una versión más sofisticada de este argumento en su famoso experimento mental de la «habitación china». No he usado ese ejemplo aquí porque el argumento de Searle está enfocado principalmente a la inteligencia (o a la «comprensión»), más que a la conciencia (Searle, 1980).

14.

El filósofo John Perry dijo en una ocasión: «Si te dedicas a pensar en la conciencia durante el tiempo suficiente, o bien te conviertes en pampsiquista o bien te metes a administrativo». Puede que el problema radique precisamente en limitarse a pensar en vez de hacer (ciencia). Como defensa elocuente del pampsiquismo, yo recomiendo

Galileo's Error, de Philip Goff (2019). La cita de Perry figura en un artículo de 2018 de Olivia Goldhill en la revista *Quartz,* en <**qz.com/1184574/the-idea-that-everything-from-spoons-to-stones-are-conscious-is-gaining-academic-credibility**>.

15.

McGinn (1989).

16.

Como argumento contrario a la idea de que hay cosas que siempre escaparán a la comprensión humana, véase Deutsch (2012).

17.

Hay un «ismo» más que también merece mención, aunque no está tan formalmente constituido como los que he descrito en el texto principal. Me refiero al *ilusionismo,* que es la idea según la cual la conciencia (fenoménica) es una ilusión introspectiva: al hacer introspección de los estados conscientes, los representamos inadecuadamente como si tuvieran unas propiedades fenoménicas *(qualia)* que, en realidad, no poseen. Según cierta interpretación –con la que no estoy de acuerdo–, el ilusionismo vendría a decir que los estados conscientes no existen en realidad. Según otra –con la que simpatizo más–, el ilusionismo dice que las experiencias conscientes existen, pero no son lo que pensamos que son. Es posible, aunque no seguro, que esta interpretación del ilusionismo sea compatible con lo que digo en este libro. Véase más sobre el ilusionismo en Frankish (2017).

18.

Existen, al menos, dos tipos de zombi filosófico. Los «zombis conductuales» son indistinguibles por fuera (en cuanto a sus comportamientos) de sus equivalentes conscientes. Los «zombis neurológicos» aportan la fantasía adicional de ser idénticos a una criatura consciente también por dentro. Un zombi neurológico posee la misma estructura interna y podría construirse también a partir del mismo *wetware* electroquímico que la criatura consciente de la que es una versión zombi. Aunque, eso sí, todas las variedades de zombi carecen por completo de conciencia.

19.

Gidon *et al.* (2020); Herculano-Houzel (2009).

20.

Los defensores del argumento de los zombis pueden responder a esto que lo único que importa es su posibilidad lógica, y no su concebibilidad en función de las leyes de la física vigentes en este universo particular. Yo discrepo. Un A380 volando hacia atrás podría ser lógicamente posible si rigieran unos principios aerodinámicos alternativos, pero aceptar esa posibilidad no arroja luz alguna sobre cómo un A380

de verdad vuela en realidad, en este mundo, con las leyes de la física y la aerodinámica que tenemos realmente. Lo que quiero saber es cómo un cerebro auténtico (y un cuerpo, y así sucesivamente), regido por las leyes de la física en *este* universo, modula y origina realmente la experiencia consciente en *este* universo.

21.

Esta perspectiva del «problema real» no es ninguna novedad, o por lo menos, no lo es del todo. El propio Chalmers describió estrategias similares cuando habló del «problema de la "puesta en correspondencia"» (o *mapping*, Chalmers, 1996), y también del «principio de coherencia estructural» (Chalmers, 1995a), integrado en su descripción original del problema difícil. Existe asimismo un abundante y muy influyente conjunto de trabajos sobre «neurofenomenología» que tratan de poner en correspondencia las propiedades fenomenológicas con aspectos del cerebro y de su actividad (Thompson, 2014; Varela, 1996). No obstante, hay diferencias de énfasis entre esas posturas (Seth, 2009, 2016b).

22.

Sutherland (1989).

23.

Crick y Koch (1990). Más o menos por la misma época, el filósofo estadounidense Daniel C. Dennett publicó su influyente *La conciencia explicada* (1991). Leer ese libro a comienzos de los años noventa supuso para mí un punto de inflexión. Sigue siendo una lectura desafiante y reveladora. Véase más sobre la historia de la ciencia de la conciencia en LeDoux *et al.* (2020) y Seth (2017, 2018).

24.

Crick y Koch (1990).

25.

Nótese que los CNC suelen interpretarse como si estuvieran relacionados con regiones cerebrales determinadas, pero no es necesario que sea así. La definición de un CNC nos refiere a un mecanismo neuronal que podría ponerse en práctica en varias regiones del cerebro. Los circuitos cerebrales implicados en un CNC dado incluso podrían cambiar con el tiempo (G. M. Edelman y Gally, 2001).

26.

Un matiz en este caso es que las regiones cerebrales que se revelan en los estudios con rivalidad binocular suelen estar asociadas a *transiciones* entre percepciones conscientes, y esto es importante porque la percepción de cambio no es lo mismo que el cambio de percepción. Véase Blake *et al.* (2014). Volveré sobre esta distinción en el capítulo 6.

27.

Hacía ya años que se hablaba de estos problemas, pero, por lo general, se les restaba importancia o incluso se tapaban barriéndolos figurativamente bajo la alfombra. Dos artículos de 2012 ayudaron por fin a concretar la cuestión: Aru *et al.* (2012), y De Graaf *et al.* (2012).

28.

Algunos experimentos han hecho el valiente intento de distinguir entre la percepción consciente, por un lado, y la atención o el acto conductual de informar, por el otro. Los resultados de los llamados paradigmas «sin informe», en los que los voluntarios no efectúan informes conductuales de lo que perciben, son particularmente interesantes. En muchos de estos estudios, el CNC restante no implica a las regiones frontales del cerebro que son centrales en teorías como la del espacio de trabajo global o la del pensamiento de orden superior. Véanse análisis recientes del tema en Frässle *et al.* (2014); Tsuchiya *et al.* (2015), y Raccah *et al.* (2021).

29.

Otra forma de expresar esta ambición es la de los filósofos Susan Hurley y Alva Noë, que distinguen entre las brechas explicativas «comparativas», relativas a la explicación de por qué las diferentes experiencias poseen las propiedades fenomenológicas concretas que poseen, y la brecha explicativa «absoluta», que es el problema (difícil) de por qué y cómo es que existe eso que llamamos fenomenología. De ese modo, podríamos concebir el problema real como un esfuerzo exhaustivo por ir cerrando las brechas explicativas comparativas a fin de resolver –y quizá disolver– en última instancia la brecha explicativa absoluta. Véase Hurley y Noë (2003).

30.

El vitalismo sostiene que «los organismos vivos son fundamentalmente diferentes de los entes no vivos porque contienen cierto elemento no físico que no está presente en las cosas inanimadas o se rigen por principios diferentes de estas últimas» (Bechtel y Williamson, 1998). Todavía hoy una mayoría de los niños y niñas en edad preescolar tienden a preferir las explicaciones vitalistas de la vida que otras más modernas (Inagaki y Hatano, 2004). Los paralelismos históricos entre el vitalismo y la ciencia de la conciencia han sido objeto de un estudio particularmente meticuloso por parte de la filósofa Patricia Churchland (Churchland, 1996).

31.

Puede que ese anhelo de una solución o momento «eureka» explique en parte el persistente atractivo de las teorías de la conciencia basadas en la mecánica cuántica, la mayoría de las cuales se remontan al libro del matemático Roger Penrose, *La nueva mente del emperador,* publicado originalmente en 1989. Aunque no se puede descartar que alguna futura teoría de base cuántica tenga algo relevante que decir

acerca de la conciencia, los intentos desarrollados hasta el momento solo me han parecido muestras de un falso silogismo: la mecánica cuántica es misteriosa; la conciencia es misteriosa; así pues, deben de estar relacionadas.

2. MEDIR LA CONCIENCIA

1.

Véase <**www.encyclopedia.com/science/dictionaries-thesauruses-pictures-and-press-releases/dalence-joachim**>.

2.

H. Chang (2004).

3.

Posiblemente, la primera persona que se hizo esta pregunta y la contestó fue Thomas Henry Huxley (conocido por el sobrenombre del «Bulldog de Darwin»). En una conferencia que impartió en 1870, dijo: «Creo que, tarde o temprano, alcanzaremos a establecer un equivalente mecánico de la conciencia del mismo modo que hemos alcanzado a establecer un equivalente mecánico del calor». Citado en Cobb (2020), p. 113.

4.

Seth *et al.* (2008).

5.

Weiser *et al.* (2008).

6.

Miles *et al.* (2004).

7.

Nasraway *et al.* (2002).

8.

Herculano-Houzel (2009).

9.

Véase Lemon y Edgley (2010). El cerebelo suele aparecer como un «área de interés» en los estudios con neuroimagen. A menudo, esos hallazgos no se comentan porque el cerebelo no era el foco de atención del experimento, por lo que, muchas veces, los investigadores no saben qué decir al respecto.

10.

DiNuzzo y Nedergaard (2017).

11.

Ferrarelli *et al.* (2010); Massimini *et al.* (2005).

12.

Para crear una representación comprimida de una secuencia solo de unos (o solo de ceros) con una longitud *n,* lo único que necesitas es especificar el uno (o el cero) inicial y, a continuación, las *n* repeticiones. En el extremo opuesto, el de una secuencia absolutamente aleatoria, no existe posibilidad de compresión alguna: tienes que especificar cada uno y cada cero en su lugar correcto para poder regenerarla luego. Las secuencias que contienen cierta dosis de estructura predecible tendrán algunas representaciones comprimidas por en medio. La longitud de la representación comprimida de forma óptima recibe el nombre de complejidad Kolmogorov-Chaitin-Solomonoff. La complejidad LZW se aproxima a (y proporciona un límite superior para) esa cantidad.

13.

Casali *et al.* (2013).

14.

Schartner *et al.* (2017b); Schartner *et al.* (2015). En esos estudios, nos aseguramos de que los cambios observados en la complejidad LZW no fuesen efectos triviales de otros cambios (más conocidos) en las señales cerebrales de EEG que acompañan a la pérdida de conciencia, como, por ejemplo, los incrementos en la potencia de las ondas «delta» de baja frecuencia que se sabe que se producen mientras dormimos.

15.

Casarotto *et al.* (2016).

16.

Véase, por ejemplo, la muy utilizada escala de coma de Glasgow (Teasdale y Murray, 2000).

17.

En la mayoría de casos, el síndrome de enclaustramiento es causado por un daño en el puente troncoencefálico, una parte del tallo cerebral que contiene importantes vías neuronales que conectan la médula espinal con otras partes del cerebro.

18.

Bauby (1997). Tal vez pienses que estar «enclaustrado» es peor que morirse, pero si bien es cierto que a algunas personas que sufren este síndrome la vida les resulta insoportable, hay un sorprendente porcentaje de ellas que manifiestan tener una calidad de vida razonablemente buena (Rousseau *et al.,* 2015), lo cual no deja de abundar en los peligros de asumir cómo son las cosas por dentro a partir de lo que

parecen por fuera.

19.

Esta historia se cuenta también aquí: <**www.humanbrainproject.eu/en/follow-hbp/news/measures-of-consciousness-and-a-tale-of-cultural-loss**>.

20.

Adrian Owen relata la historia del descubrimiento que hizo su equipo, así como las implicaciones que puede tener para la medicina y para la investigación de la conciencia, en su libro *Into the Gray Zone* (Owen, 2017).

21.

Monti *et al.* (2010). Dado que los pacientes solo pueden responder «sí» o «no» tal como está diseñada la prueba, los investigadores tienen que andarse con mucho ojo a la hora de seleccionar sus preguntas. Preguntarles si sienten dolor, por ejemplo, es razonable, pero ¿y preguntarles si quieren seguir viviendo? Hay ahí muchos dilemas éticos implicados. En 2016, la escritora Linda Marshall Griffiths, la productora Nadia Molinari y yo creamos un *audiodrama* (una obra de teatro en audio) titulado *The Sky Is Wider* en el que se exploran esas cuestiones, en <**https://www.lindamarshallgriffiths.co.uk/the-sky-is-wider-best-single-drama**>.

22.

Naci *et al.* (2017).

23.

Un ejemplo más famoso de este mismo problema es el intento de definir y medir la inteligencia con respecto a una escala unitaria como el CI. Este proyecto siempre ha chocado contra la realidad de la gran variedad de comportamientos inteligentes observados en diferentes individuos, culturas y especies. Véase en Baine *et al.* (2016) un método multidimensional de caracterización del nivel consciente.

24.

Véase Konkoly *et al.* (2021).

25.

Albert Hoffmann, «LSD–My Problem Child», <**https://psychedelic-library.org/child1.htm**>.

26.

Véase más sobre esta historia en Michael Pollan, *Cómo cambiar tu mente* (Pollan, 2018).

27.

Carhart-Harris *et al.* (2012).

28.

Los datos de MEG y de EEG suelen ser de muy alta resolución temporal, del orden de milisegundos, pues reflejan directamente la actividad eléctrica de poblaciones enteras de neuronas. La IRMf es muy lenta en comparación, pues tiene una escala temporal natural de segundos. Esto se debe, en parte, a la lentitud con la que funciona un escáner de IRM típico, pues normalmente solo permite una medición cada uno o dos segundos, y, en parte también, a la lentitud de la escala temporal de la señal de oxigenación de la sangre que mide el IRMf.

29.

Schartner *et al.* (2017a). En otro estudio, más reciente, se ha detectado el mismo patrón de resultados con la DMT (Timmermann *et al.,* 2019).

30.

En un estudio de seguimiento dirigido por Lionel Barnett, hemos descubierto que el estado psicodélico implica unas reducciones significativas del «flujo de información» entre regiones de la corteza cerebral. Esto de nuevo evidencia una pérdida de estructura perceptiva durante el estado psicodélico. Véase Barnett *et al.* (2020).

31.

Tononi y Edelman (1998). Edelman falleció en 2014; yo mismo escribí uno de los numerosos obituarios que rindieron homenaje a tan influyente científico (Seth, 2014a).

32.

Algunos filósofos han cuestionado el supuesto de que las experiencias conscientes sean necesariamente unos todos unificados (Bayne, 2010). Una situación en la que a menudo se plantea la posibilidad de una conciencia no unificada es la de los pacientes con «cerebro dividido», a los que se les han separado quirúrgicamente los hemisferios corticales. Véase más al respecto en el capítulo 3.

33.

Tononi *et al.* (1994).

34.

Seth *et al.* (2011a); Seth *et al.* (2006).

35.

Para una comparación entre indicadores de complejidad en el contexto de la neurociencia, véanse Seth *et al.* (2011a) y Mediano *et al.* (2019).

36.

Demertzi *et al.* (2019); Luppi *et al.* (2019).

3. FI

1.

Algunos de los artículos principales sobre la TII escritos por miembros del grupo de Tononi son: Tononi (2008); Tononi (2012); Tononi *et al.* (2016). Véase un manifiesto de la teoría muy accesible en Koch (2019).

2.

Seth *et al.* (2006).

3.

Véase <**www.scientificamerican.com/article/is-consciousness-universal**>.

4.

Véanse, por ejemplo, las críticas de Scott Aaronson en <**www.scottaaronson.com/blog/?p=1799**>.

5.

Φ también se puede concebir como una forma de medir la propiedad de la «emergencia». La emergencia es un concepto muy general que alude a cómo las propiedades macroscópicas (como una bandada) surgen de –o se relacionan con– sus componentes microscópicos (las aves individuales, en este caso). Véanse Hoel *et al.* (2013); Rosas *et al.* (2020); Seth (2010).

6.

Los propios Tononi y Koch usan esa analogía (Tononi y Koch, 2015).

7.

Los tomo adaptados de Tononi (2008).

8.

Hablamos de una primera aproximación, dejando a un lado factores como el ajuste del contraste, cuyo funcionamiento sí puede implicar a todo el conjunto.

9.

Las operaciones de división del cerebro consisten en seccionar el voluminoso haz de fibras nerviosas que interconectan los hemisferios corticales: el cuerpo calloso. La cirugía puede ir bien para aliviar la epilepsia grave, pero hoy en día es muy poco frecuente, dada la disponibilidad de cada vez más opciones de tratamiento alternativas menos invasivas. Los pacientes con cerebro dividido conservan cierto grado de conectividad entre los hemisferios, pero a los efectos de este ejemplo, supongamos que la totalidad de su cerebro ha sido cortada por la mitad. Véase De Haan *et al.* (2020).

10.

Tononi *et al.* (2016).

11.

Además de los axiomas sobre la integración y la información, la TII postula otros tres: que la conciencia existe, que se compone de múltiples elementos, y que es exclusiva de una escala espaciotemporal particular (Tononi *et al.,* 2016). El filósofo Tim Bayne ha criticado la TII basándose en que los axiomas que propone, sobre todo el último (el referido a la «exclusión»), tal vez no sean tan evidentes (Bayne, 2018).

12.

En teoría de la información, la información en sí (que es reducción de incertidumbre) se mide utilizando una cantidad llamada entropía. La entropía (cuyo símbolo habitual es la letra S) es una función del número de estados distintos en los que puede estar un sistema y de la probabilidad de que esté en cada uno de esos estados. Viene dada por la fórmula $S = -\Sigma p_k \log(p_k)$. Dicho en palabras, esto significa que para cada estado *(k)* de un sistema, se multiplica la probabilidad de estar en ese estado por el logaritmo de esa probabilidad, y luego se suman esos valores para todos los estados. Para cualquier sistema dado, la entropía será máxima cuando cada estado sea igualmente probable. Un dado de seis caras justo posee una entropía de unos 2,5 bits (cuando el logaritmo se calcula en base 2). Un dado cargado tendrá una entropía más baja.

13.

Así pues, el valor Φ de un sistema dice más sobre su mecanismo (cómo está «cableado») que sobre su dinámica (qué hace). De hecho, algunas versiones recientes de la TII describen Φ en términos de «poder de causa-efecto irreductible», es decir, en términos de mecanismo, no de dinámica (Tononi *et al.,* 2016).

14.

Técnicamente hablando, esta división se denomina la *partición de información mínima.* Hay también problemas relacionados con cuál puede ser el mejor modo de tratar particiones de diferentes tamaños, dado que cuanto mayor sea una partición, al tener más elementos, más información será capaz de generar.

15.

Según la TII, podría haber un «país consciente» si se maximiza Φ para una escala espaciotemporal que abarque un país entero, tal vez concibiendo a las personas individuales como elementos análogos a las neuronas en un cerebro. Una situación así tendría la peculiar implicación de que, desde el momento en que consideráramos que un país es consciente, sus elementos individuales –las personas– dejarían de tener conciencia individual. Esta rocambolesca posibilidad ha sido estudiada por el filósofo estadounidense Eric Schwitzgebel,

<schwitzsplinters.blogspot.com/2012/03/why-tononi-should-think-that-united.html>.

16.

Deisseroth (2015).

17.

Véase <www.templetonworldcharity.org/accelerating-research-consciousness-our-structured-adversarial-collaboration-projects>. El experimento «inactivas versus inactivadas» fue una propuesta de Umberto Olcese y Giulio Tononi.

18.

Wheeler (1989).

19.

Barrett y Seth (2011); Mediano *et al.* (2019). Desde el punto de vista matemático, nuestros indicadores funcionan con la distribución empírica de un sistema, más que con su distribución de máxima entropía.

20.

Una propuesta interesante en ese sentido es la de que la «espacialidad» dominante de la experiencia visual se explica por la anatomía en cuadrícula descubierta en niveles inferiores del córtex visual (Haun y Tononi, 2019).

4. PERCIBIR DE DENTRO HACIA FUERA

1.

Las jerarquías perceptivas no están estrictamente aisladas entre sí. Una buena regla general en ese sentido es aquella que dice que, cuanto más nos alejamos de la periferia sensorial, más interacción transmodal apreciamos. Véanse Felleman y Van Essen (1991); Stein y Meredith (1993).

2.

Grill-Spector y Malach (2004).

3.

Marr (1982).

4.

He *et al.* (2016).

5.

Anscombe (1959), p. 151, la cursiva es la del original.

6.

Tomado del *Libro de la óptica,* de Alhacén (de hacia el año 1030), traducido al inglés en Sabra (1989). La historia se cuenta con mayor detalle en el estupendo libro *The Predictive Mind,* de Jakob Hohwy (2013).

7.

Swanson (2016).

8.

Bruner y Goodman (1947).

9.

Gregory (1980).

10.

Clark (2013); Clark (2016); Hohwy (2013); Rao y Ballard (1999).

11.

Seguramente, no todas las predicciones proceden de arriba abajo ni todos los errores de predicción van necesariamente de abajo arriba. Se pueden concebir unas «predicciones ascendentes» que funcionen como restricciones que reflejen aspectos globales y estables de la inferencia perceptiva (Teufel y Fletcher, 2020). Un ejemplo de tal restricción perceptiva podría ser la sobrerrepresentación de orientaciones verticales y horizontales en las imágenes naturales. La influencia de las sombras sobre la luminancia –de la que hablaremos más adelante– tal vez constituya otra de esas restricciones predictivas ascendentes.

12.

Véase Dennett (1998). He aquí otro ejemplo: para escuchar música, no necesitamos una banda diminuta en la cabeza ni un intrincado sistema de micrófonos intracraneales. Daniel Dennett también es el autor del feliz juego de palabras entre los términos ingleses *pigment* («pigmento») y *figment* («figuración») que creó con la expresión *mental pigment* aplicada al caso de la percepción del color.

13.

Gasquet (1991). Esta es una cita que también se ha atribuido a Paul Klee.

14.

Muchas gracias al artista y rapero Baba Brinkman (que tiene la peculiaridad de someter sus creaciones a un proceso de revisión por pares) por esta provocadora frase sobre la realidad. En el capítulo 6, hablaremos más sobre la percepción de «lo que es real».

15.

En un estudio reciente, Biyu He y un equipo de colaboradores han examinado cómo difieren las dinámicas neuronales cuando se identifica una imagen en un test de Mooney respecto a cuando no. Véase Flounders *et al.* (2019).

16.

Brainard y Hurlbert (2015); Witzel *et al.* (2017).

17.

Eye Benders (Gifford y Seth, 2013).

18.

Chris Darwin tiene publicados en línea unos excelentes ejemplos de habla sinusoidal, en <**www.lifesci.sussex.ac.uk/home/Chris_Darwin/SWS**>. Yo mismo utilizo otro ejemplo en mi charla TED de 2017: <**www.ted.com/talks/anil_seth_your_brain_hallucinates_your_conscious_reality**>. También hay equivalentes auditivos del fenómeno del vestido. Un ejemplo es un sonido en el que unas personas advierten el nombre «Yanny» y otras el nombre «Laurel» (Pressnitzer *et al.,* 2018). En 2020, apareció un vídeo en TikTok en el que parece que el ruidito emitido por un juguete barato suena como *green needle* («aguja verde») o *brainstorm* («tormenta de ideas») dependiendo de las palabras que estés leyendo en ese momento (<**time.com/5873627/green-needle-brainstorm-explained**>).

19.

Véase en De Lange *et al.* (2018) un panorama general de los experimentos que muestran cómo las expectativas influyen en la percepción.

20.

Otro caso muy útil para entender intuitivamente mejor la cuestión de la veracidad perceptiva es el de la visión periférica: esa parte del campo visual alejada del centro de tu mirada (la «visión foveal»). La periferia visual tiene una densidad de fotorreceptores mucho menor que la fóvea, y sin embargo, la experiencia visual en esa periferia no nos resulta borrosa. ¿Significa eso que la visión periférica es menos verídica que la foveal porque manifiesta una «ilusión de claridad» (o, para ser más específicos, una ilusión de «no borrosidad»)? ¡No! La claridad y la borrosidad son propiedades de la experiencia perceptiva relativas a los datos sensoriales a los que está ajustada cada parte del sistema visual. Véase un esclarecedor análisis en Haun (2021), un poco de contexto histórico en Lettvin (1976) y un estudio de largo alcance sobre la (no) veracidad perceptiva en Hofman *et al.* (2015).

21.

John Locke, *Ensayo sobre el entendimiento humano* (1689), 14.ª edición inglesa (1753). Aunque el ejemplo del color es intuitivo, los filósofos han debatido con denuedo si este puede considerarse realmente una cualidad secundaria (Byrne y Hilbert, 2011).

Una distinción relacionada con esta que aparece también en la literatura filosófica es la que se establece entre diferentes géneros de «géneros», en la que las diferencias radican en las condiciones necesarias para que algo exista. Por ejemplo, el dinero requiere de unas convenciones sociales para existir, por lo que es un «género social». El agua no precisa de tales convenciones sociales para su existencia, por lo que es un «género natural».

5. LA MAGIA DE LAS PROBABILIDADES

1.

En el original inglés, el título de este capítulo era «The Wizard of Odds», un juego con la palabra *odds* (probabilidades) y el título del famoso libro (y de la célebre película homónima) *The Wizard of Oz [El mago de Oz]*. Doy nuevamente las gracias a Baba Brinkman, inspirador del mencionado título, que está tomado de su *Rap Guide to Consciousness* (2018), obra sonora de la que fui asesor científico. Véase <**https://bababrinkman.com/shows/#consciousness**>.

2.

Este ejemplo es una adaptación de otro de F. V. Jensen (2000). Para adquirir una muy completa base sobre razonamiento abductivo, no hay mejor guía a la que recurrir que *Inference to the Best Explanation,* de Peter Lipton (2004).

3.

Así es como se formula normalmente el teorema de Bayes:

$$p(H|D) = \frac{p(D|H) * p(H)}{p(D)}$$

$p(H|D)$ es la probabilidad *a posteriori* de la hipótesis H dados los datos D; $p(D|H)$ es la probabilidad de los datos dada la hipótesis, y $p(H)$ y $p(D)$ son las probabilidades *a priori* de la hipótesis y de los datos, respectivamente. Cabe señalar que puede ser difícil calcular un valor para $p(D)$, pero, por suerte, suele no ser necesario. Cuando el objetivo es hallar la hipótesis más probable *a posteriori* entre una serie de opciones alternativas (como suele ser el caso), las $p(D)$ de todas ellas se suelen anular mutuamente.

4.

En 2009, el Gobierno estadounidense aconsejó no realizar exámenes mamográficos masivos para la detección precoz del cáncer de mama en mujeres de entre cuarenta y cincuenta años justamente por esa razón. Las mamografías tenían en aquel entonces una sensibilidad de aproximadamente el 80 %, lo que significa que detectaban en

torno a un 80 % de los casos de mujeres de ese grupo de edad que tenían realmente cáncer de mama y se sometían al examen. Pero las pruebas también daban positivos de cáncer en aproximadamente un 10 % de las mujeres que, en realidad, no tenían la enfermedad en ese momento. El factor crucial aquí era que el cáncer de mama presenta una incidencia –una tasa «base» o «de referencia»– relativamente baja, de aproximadamente el 0,04 % para todo ese grupo de edad. Aplicando el teorema de Bayes y usando esa incidencia de referencia como probabilidad *a priori,* podemos calcular que la probabilidad de tener cáncer de mama en la hipótesis de contar con un positivo en una mamografía es de solo un 3 %. De cada cien mujeres que den positivo en el examen, noventa y siete estarán libres de cáncer en realidad y pasarán por un periodo de innecesaria ansiedad, así como por posibles pruebas y estudios adicionales potencialmente caros e invasivos. Una de las conclusiones que se extrajo de toda esa historia fue la necesidad de mejorar la sensibilidad y la especificidad de las pruebas, y de hecho, las mamografías son hoy mucho mejores que entonces. Un estudio reciente en el Reino Unido ha sugerido que ahora sí puede valer la pena ya someter a ese tipo de revisiones a todas las mujeres de entre cuarenta y cincuenta años de edad. Véanse McGrayne (2012); Duffy *et al.* (2020).

5.

Sharon McGrayne cuenta magníficamente bien la historia –sorprendentemente controvertida– del análisis bayesiano en su *La teoría que nunca murió* (2012).

6.

Véase más al respecto en Lakatos (1978) y en Seth (2015b).

7.

Todo mi reconocimiento para Paul Fletcher y Chris Frith, quienes argumentaron esto mismo a propósito de su teoría bayesiana de la alucinación y el delirio en la esquizofrenia (Fletcher y Frith, 2009).

8.

Estrictamente hablando, X es una *variable aleatoria* porque su valor está determinado por una distribución de probabilidad. El ejemplo mostrado en la figura de la p. 103 es una distribución *continua* de probabilidad (también llamada función de densidad de probabilidad), ya que X puede adoptar cualquier valor dentro del intervalo permitido. Si X solo pudiera tomar valores concretos –como, por ejemplo, «cara» o «cruz»–, hablaríamos de una distribución *discreta* de probabilidad.

9.

Hay dos maneras de interpretar los mapas de correspondencia entre las creencias bayesianas y el cerebro. La versión débil dice que nosotros, como observadores externos, usamos esas creencias para representarnos cosas a *nosotros* mismos, igual

que podríamos utilizar un mapa físico para representar nuestro entorno. Según esa perspectiva, la actividad neuronal detectada podría representar ciertas situaciones observadas por nosotros como científicos. La interpretación fuerte es que el cerebro emplea esas creencias (o algo que se les parece) para representarse cosas para *sí mismo.* Esta segunda interpretación ejemplifica la hipótesis del «cerebro bayesiano» y es fundamental para el concepto del cerebro como máquina de predicción. No haber distinguido bien entre esas dos acepciones de «representación» ha sido el motivo de no poca confusión en muchos ámbitos de la ciencia cognitiva y de la neurociencia. Véase Harvey (2008).

10.

Hay quienes se refieren a la vía ascendente *(bottom-up)* y a la descendente *(top-down)* con los términos «proalimentación» y «retroalimentación», respectivamente. Pero, desde la perspectiva del procesamiento predictivo, eso significa entenderlo todo justo al revés. En ingeniería, la «retroalimentación» suele relacionarse con una señal de error que se usa para ajustar la señal de control «de proalimentación». Por consiguiente, en el procesamiento predictivo, las conexiones ascendentes pasan a ser conexiones «de retroalimentación», porque son las que transmiten las señales de error. Para complicar aún más las cosas, como ya se ha mencionado anteriormente, algunas predicciones (globales, estables) pueden ir inscritas en señales que proceden de abajo arriba, en cuyo caso, las conexiones que van de arriba abajo pasan a transmitir errores de predicción (Teufel y Fletcher, 2020).

11.

En este libro, uso el concepto de procesamiento predictivo como una manera abreviada de referirme a una serie de teorías diversas, entre las que se incluye la codificación predictiva, que incorpora el mecanismo fundamental de la minimización del error de predicción. Con esto no he pretendido minimizar sus diferencias, que son tan interesantes como importantes (Hohwy, 2020a).

12.

La minimización jerárquica del error de predicción indica que las partes del cerebro implicadas en la percepción deben de estar más dotadas de esa conectividad descendente que transmite señales desde los niveles jerárquicos superiores hacia los inferiores. Muchos estudios así lo han demostrado; véase, por ejemplo, Markov *et al.* (2014). La presencia de esa rica conectividad descendente resulta más difícil de explicar desde una perspectiva de la percepción entendida como un proceso que procede «de abajo arriba».

13.

Véase Feldman y Friston (2010). La ponderación de la precisión se logra modificando las creencias *a priori* relativas a las precisiones (también llamadas *hiper-a prioris)* para

incrementar o disminuir la precisión inferida.

14.

Véase también Simons y Chabris (1999).

15.

Para un estudio general sobre psicología y magia, véase Kuhn *et al.* (2008).

16.

Mi colega Andy Clark, de la Universidad de Sussex, lleva tiempo defendiendo estas formulaciones del procesamiento predictivo «orientado a la acción». Su libro *Surfing Uncertainty* (2016) es un recurso fundamental en este terreno.

17.

Como sostiene el neurocientífico György Buzsáki en su libro *The Brain from Inside Out,* esta perspectiva plantea retos –a la vez que abre nuevas oportunidades– para la neurociencia experimental. La mayoría (aunque, ni mucho menos, la totalidad) de los experimentalistas estudian el cerebro examinando su actividad en respuesta a una estimulación externa, y no entendiéndolo como un sistema intrínsecamente dinámico y activo. Véase Buzsáki (2019), y también Brembs (2020).

18.

Friston *et al.* (2010).

19.

Alexander Tschantz, Christopher Buckley, Beren Millidge y yo llevamos ya un tiempo desarrollando sobre esta base nuevos algoritmos de aprendizaje automático que son capaces de aprender modelos generativos a partir de pequeñas cantidades de datos (Tschantz *et al.,* 2020a). Curiosamente, la posibilidad de elaborar predicciones ascendentes (Teufel y Fletcher, 2020) tiene una prometedora aplicación en este contexto. Se puede relacionar con la potente técnica de la «amortización» en el aprendizaje automático, donde una creencia bayesiana *a posteriori* aproximada se calcula por medio del barrido en proalimentación (descendente) de una red neuronal artificial entrenada a ese efecto.

20.

Esta idea se puede contrastar experimentalmente. Y, de hecho, resulta que, durante la acción, la sensibilidad sensorial propioceptiva se reduce, tal como cabría esperar (C. E. Palmer *et al.,* 2016). La atenuación sensorial que acompaña a la acción también sirve de muy buena explicación de por qué no podemos hacernos cosquillas a nosotros mismos (H. Brown *et al.,* 2013).

6. LA PARTICIPACIÓN DEL ESPECTADOR

1.

Kandel (2012).

2.

Gombrich (1961).

3.

Seth (2019b).

4.

Kandel (2012), p. 204.

5.

Este ejemplo está tomado de Albright (2012) y tal como se adaptó en Seth (2019b). Véase <**www.wikiart.org/en/camille-pissarro/hoarfrost-1873**>.

6.

La expresión de Leroy apareció en una reseña satírica sobre el arte impresionista publicada en *Le Charivari* el 25 de abril de 1874, que fue precisamente la reseña en la que se acuñó el término «impresionismo».

7.

Gombrich (1961).

8.

En sus años finales, Pissarro padeció graves problemas oculares. Sus compañeros impresionistas Claude Monet y Edgar Degas sufrieron dolencias parecidas. Sería interesante saber si (y de qué modo) sus problemas de vista contribuyeron en algo a su intuición artística. Aunque no sería descartable que influyeran un poco –quizás eran más sensibles a los patrones lumínicos que a los detalles de los objetos–, el caso de otro conocido pintor, El Greco (1541-1614), nos invita a ser cautos en ese sentido. Las obras de El Greco solían incluir figuras antinaturalmente alargadas, un rasgo que algunos atribuyeron a su pronunciado astigmatismo. Según esa historia, él pintaba figuras alargadas porque así veía la realidad. Pero los psicólogos ya se han encargado de señalar (muy correctamente, por cierto) que, según esa lógica, el pintor también habría visto su lienzo más alargado de lo normal, y eso habría compensado el presunto efecto de su astigmatismo. A esta clase de error lógico se le ha llegado a llamar incluso la «falacia de El Greco», una falacia en la que aún hoy continúan cayendo algunos investigadores de la percepción. Véase Firestone (2013).

9.

Gombrich (1961), p. 170.

10.

De esta idea de Gombrich se haría posteriormente eco el escritor, crítico y artista John Berger, cuyo influyente libro de 1972 *Modos de ver* comenzaba con la frase: «Nunca se ha establecido la relación entre lo que vemos y lo que sabemos». Frente a un conservador cultural como Gombrich, Berger quiso enfatizar las influencias políticas y culturales que condicionan la percepción y poner de relieve que lo que vemos puede ser diferente según las personas y según los grupos (Berger, 1972).

11.

En realidad, esta predicción no se deduce tan directamente como parece; véase Press *et al.* (2020).

12.

Esta es una variante de otro método más conocido que describí en el capítulo 1: el de la «rivalidad binocular» (Blake *et al.,* 2014).

13.

Pinto *et al.* (2015). Nuestro estudio fue más complejo de lo que este resumen puede dar a entender. Llevamos a cabo un gran número de estudios de control para descartar al máximo otros factores que pudieran explicar nuestros resultados, como ciertos sesgos de las personas en cuanto a cómo reaccionan o a cómo centran su atención. Para un repaso general de otros estudios similares, véase De Lange *et al.* (2018); véase también una influyente aportación temprana a esta literatura especializada en Melloni *et al.* (2011).

14.

La «lectura cerebral» consiste en entrenar a unos algoritmos de aprendizaje automático para que clasifiquen la actividad cerebral en diferentes categorías. Véase Heilbron *et al.* (2020).

15.

Imagen tomada de <**www.boredpanda.com/objects-with-faces**>.

16.

Suzuki *et al.* (2017).

17.

En concreto, hablamos de redes neuronales convolucionales profundas (RNCP) que se pueden entrenar mediante algoritmos de retropropagación estándar. Véase Richards *et al.* (2019).

18.

En el modo convencional («hacia delante»), se le muestra una imagen a la red, la actividad se propaga hacia arriba a través de las diversas capas y el *output* de la red

nos dice qué «piensa» que hay en la imagen. En el algoritmo Deep Dream (y en la adaptación que de este ha hecho Keisuke), se invierte ese proceso. Fijamos el *output* de la red y ajustamos el *input* hasta que la red alcanza un estado estable. Véanse los detalles en Suzuki *et al.* (2017).

19.

La experiencia completa con la máquina de alucinaciones –en la que se puede ver un vídeo panorámico a través de unas gafas de realidad virtual– es mucho más inmersiva que un simple fotograma o una instantánea. Hay una grabación de muestra de aquella película en <**www.youtube.com/watch?v=TlMBnCrZZYY**>.

20.

Cuando se hacen corresponder esos modelos informáticos con unas hipótesis sobre circuitos neurales, se emplea un enfoque llamado «neurofenomenología informática»: una versión potenciada por ordenador de la neurofenomenología cuyos orígenes se remontan a Francisco Varela (1996). Con esa misma voluntad, Keisuke Suzuki, David Schwartzman y yo estamos desarrollando nuevas versiones de la máquina de alucinaciones que incorporen explícitamente modelos generativos, y que, por lo tanto, se puedan hacer corresponder de forma más ajustada con lo que creemos que está ocurriendo en los cerebros reales. Nuestras nuevas máquinas de alucinaciones pueden captar una variedad mucho mayor de experiencias alucinatorias que la original.

21.

Seth (2019b).

22.

Merleau-Ponty (1964).

23.

Véanse O'Regan (2011); O'Regan y Noë (2001). Como todas las teorías, esta se basa en una gran cantidad de trabajo previo, entre el que destacan, en su caso, las ideas de James Gibson sobre cómo la percepción depende de la acción corporeizada, y también la fenomenología filosófica de Edmund Husserl y Maurice Merleau-Ponty. La noción de «ofrecimientos» *(affordances)* de Gibson –que volveremos a ver en el capítulo 9– capta la idea de que percibimos los objetos en términos de las posibilidades para la conducta que nos «ofrecen» (Gibson, 1979). Husserl escribió que «la percepción tiene horizontes de otras posibilidades de la percepción, percepciones que *podríamos* tener si *dirigiésemos* de un modo *activo* la marcha de la percepción por otro rumbo» (Husserl, 1960 [1931]). Merleau-Ponty, que estaba muy influido por Husserl, hizo hincapié en los aspectos corporeizados de la experiencia perceptiva, y su *Fenomenología de la percepción* sigue siendo una obra influyente

(Merleau-Ponty, 1962).

24.

Este es otro ejemplo de «fenomenología informática». Véase también en Cisek (2007) una teoría neurofisiológica relacionada formulada en términos de «competencia entre ofrecimientos».

25.

Seth (2014b).

26.

Suzuki *et al.* (2019).

27.

Otro buen ejemplo nos los proporcionan los llamados efectos secundarios ópticos del movimiento, como la ilusión de la cascada de agua. Mira fijamente una catarata (o un vídeo de una catarata) durante un rato y luego aparta la vista hacia otro lado, como, por ejemplo, hacia alguna pared rocosa que haya allí cerca. Tendrás entonces la impresión de que la pared se mueve hacia arriba y, al mismo tiempo, de que está quieta en su sitio.

28.

Muchos neurocientíficos especializados en la percepción del tiempo no estarán de acuerdo con esta tesis. De hecho, la mayoría de modelos psicológicos y neurales de la percepción del tiempo presuponen algún tipo de «marcapasos» neuronal que sirve de referencia con el que comparar el tiempo físico, lo que da lugar a las percepciones de duración (Van Rijn *et al.,* 2014). Otros piensan que la percepción del tiempo depende de señales que proceden del organismo y se parecen a las de un reloj (la frecuencia cardíaca, por ejemplo) –véase Wittmann (2013)–, pero nuestra investigación pone en duda esta idea también (Suárez-Pinilla *et al.,* 2019).

29.

La subestimación de las duraciones largas y la sobreestimación de las cortas es un ejemplo de efecto de «regresión hacia la media», que se aprecia en muchas modalidades perceptivas (cuando no en todas ellas), y es un elemento característico de la inferencia bayesiana, dado que la media puede entenderse como una creencia *a priori.* En la percepción del tiempo, este efecto se conoce por el nombre de ley de Vierordt.

30.

Véase Roseboom *et al.* (2019). Más corroborada aún quedó la idea de Warrick cuando se comprobó que la correspondencia entre el modelo informático y la actuación humana era más estrecha todavía cuando se limitaban los *inputs* de la red

solamente a las partes de cada vídeo que miraba una persona.

31.

Sherman *et al.* (2020).

32.

Stetson *et al.* (2007). Obviamente, habían colocado una gran red de seguridad por debajo para frenarles la caída.

33.

Para una versión inicial de este proyecto, véase Suzuki *et al.* (2012).

34.

Phillips *et al.* (2001). Esta línea de investigación tiene también implicaciones potencialmente importantes en sociología. Si las personas experimentan sus percepciones como algo «real» y verídico, les resultará difícil aceptar que otras puedan tener experiencias perceptivas diferentes, sobre todo, si afrontan unas mismas circunstancias objetivas. Esa es una de las razones por las que hubo tanto lío y jaleo con *El vestido* (véase el capítulo 4). A la gente que lo veía de un modo sencillamente le resultaba imposible aceptar que hubiera otras maneras de verlo, precisamente porque esas personas experimentaban sus percepciones como si para ellas fueran reveladoras de una realidad objetiva. En cualquier caso, una deriva perceptiva de esta clase –una generalización como la que es propia de las cámaras de resonancia que son las redes sociales, por ejemplo– tiene múltiples implicaciones a propósito de cómo podríamos reconocer (y resolver o dar cabida a) diferencias entre individuos, grupos y culturas (Seth, 2019c).

35.

Estas citas proceden del *Tratado de la naturaleza humana* (de 1738, 1.3.14.25) y la *Investigación sobre los principios de la moral* (de 1751, apéndice 1.19), según aparecen en Kail (2007), p. 20. Yo encontré el material en Dennett (2015).

36.

Los filósofos llaman a esto «transparencia» (Metzinger, 2003b).

7. DELIRIO

1.

Collier (2012).

2.

Davis *et al.* (2017).

8. ESPÉRATE A TI MISMO

1.

Este experimento mental ha sido atribuido de forma independiente tanto al filósofo Derek Parfit como al escritor Stanisław Lem.

2.

Esta aproximación al yo ha pasado a la posteridad filosófica con el nombre de «teoría del haz».

3.

Metzinger (2003a).

4.

Este es el caso de Krista y Tatiana Hogan, de la Columbia Británica (Canadá): <**www.nytimes.com/2011/05/29/magazine/could-conjoined-twins-share-a-mind.html**>.

5.

Véase un estudio general de trastornos del yo corporeizado en Brugger y Lenggenhager (2014).

6.

Véase una deconstrucción clásica en primera persona de esta intuición en Douglas Harding, *Vivir sin cabeza* (1961).

7.

El artículo original donde se describe la ilusión de la mano de goma se publicó en 1998 (Botvinick y Cohen, 1998) y ha dado lugar a toda una pequeña «industria» de estudios de seguimiento del tema. Hay quienes han investigado si son posibles las ilusiones de tres manos o de ninguna, o si inducir la «propiedad» de una mano con un color de piel distinto al de la persona puede modificar sesgos raciales implícitos, o incluso si los ratones son susceptibles de sentir una ilusión de la «cola de goma». Véanse estados de la cuestión sobre este campo en Braun *et al.* (2018) y en Riemer *et al.* (2019).

8.

El experimento también funcionaba cuando los voluntarios veían la espalda de un cuerpo falso, generado por ordenador (un «avatar»), en vez del suyo. Véanse Ehrsson (2007); Lenggenhager *et al.* (2007).

9.

Monroe (1971).

10.

Citado en Tong (2003), p. 105.

11.

Blanke *et al.* (2004), p. 248.

12.

Blanke *et al.* (2015).

13.

Véase Brugger y Lenggenhager (2014).

14.

Petkova y Ehrsson (2008).

15.

Véase <**www.themachinetobeanother.org**>.

16.

Lush *et al.* (2020). En ese artículo, usamos el término «control fenomenológico» en vez de «hipnosis», en parte porque este último arrastra un desafortunado bagaje histórico. La idea de que las expectativas implícitas generadas por el diseño experimental pueden afectar a las experiencias y el comportamiento de los participantes es un problema ya conocido –aunque, a menudo, no lo bastante valorado– en psicología que se remonta a los primeros trabajos sobre las llamadas «características de la demanda» (Orne, 1962). Lo más importante es que no basta con comparar solamente las caricias sincronizadas con las asíncronas para controlar los efectos de las características de la demanda en la ilusión de la mano de goma, pues las personas tienen fuertes expectativas acerca de qué experimentar en estas condiciones diferentes (Lush, 2020). Véase un resumen de nuestros estudios sobre la ilusión de la mano de goma, el control fenomenológico y las características de la demanda en Seth *et al.* (2021).

17.

Además de inducir conductas y testimonios («informes») de experiencias subjetivas, la sugestión hipnótica puede generar respuestas fisiológicas y neurofisiológicas (o reprimirlas, en el caso de la analgesia hipnótica) (Barber, 1961; M. P. Jensen *et al.,* 2017; Stoelb *et al.,* 2009). Estos indicadores, más objetivos en apariencia, también podrían ser alterados (e invalidados) por la sugestionabilidad. Por suerte, también se puede ver una oportunidad en todo esto, más allá del problema evidente que representa. La sugestión hipnótica nos proporciona una potente técnica para estudiar cómo las expectativas que operan desde arriba pueden generar (o suprimir) la experiencia perceptiva. Actualmente, estamos llevando a cabo una serie de experimentos para estudiar más a fondo qué tipos de experiencia perceptiva (del yo y

del mundo) pueden ser modificados o generados por los efectos de sugestión.

18.

El psicólogo estonio-canadiense Endel Tulving llamó a este tipo de autoconciencia «conciencia autonoética» (Tulving, 1985).

19.

Estas entradas de diario están tomadas del libro de Deborah Wearing, *Forever Today* (2005).

20.

Wearing (2005), pp. 202-203.

21.

En <**www.newyorker.com/magazine/2007/09/24/the-abyss**>.

22.

La existencia de una percepción social directa, así definida, no está universalmente aceptada. Otros enfoques proponen, más bien, que nuestra conciencia de los estados mentales de otras personas se infiere a partir de su comportamiento, por una vía distinta de la percepción. Véase más al respecto en Gallagher (2008) y en C. J. Palmer *et al.* (2015).

23.

Una de las implicaciones de la inferencia activa social es que, igual que ocurre con la inferencia activa visual, los modelos generativos subyacentes llevan codificadas unas predicciones condicionales acerca de las consecuencias de las acciones. En el caso de la vista, como comentamos en el capítulo 6, esas predicciones tienen que ver con cómo cambiarían las señales sensoriales visuales si la acción fuese una o fuese otra. En ese caso, ya comenté que esas predicciones condicionales subyacen a la propiedad fenomenológica de «lo objetual». En un artículo de 2015, Colin Palmer, Jakob Hohwy y yo mismo postulamos que algo similar ocurre en la percepción social. Nuestra idea es que los estados mentales de otras personas nos parecen «reales» por cuanto nuestro cerebro codifica un rico repertorio de predicciones condicionales acerca de cómo podrían cambiar si se diera una acción u otra. Por ejemplo, una de esas predicciones podría referirse a cómo la creencia o el estado emocional de alguien cambiaría si emitiéramos un enunciado en particular (como «tráeme vino»). Esta idea dota al hecho de que tengamos una teoría de la mente de cierto sentido parecido al de una máquina de predicción, y podría servirnos de un modo útil de interpretar determinados déficits aparentes de percepción social, como los que se observan, por ejemplo, en el autismo. Véase C. J. Palmer *et al.* (2015).

24.

Los análisis neurocientíficos de la percepción social suelen poner el énfasis en las llamadas «neuronas espejo». Estas células (descubiertas por vez primera en cerebros de mono por el neurocientífico italiano Giacomo Rizzolatti y sus colegas) se activan cuando un animal actúa y observa que otro animal realiza la misma acción (Gallese *et al.,* 1996). Se dice, en ese caso, que las neuronas reflejan como un «espejo» la conducta de otros animales, porque reaccionan como si el propio observador estuviera actuando. Se ha llegado a postular que la capacidad de estas neuronas para responder de ese modo sería el fundamento sobre el que se basan toda clase de fenómenos sociales. Pero tales hipótesis depositan demasiada carga explicativa sobre lo que, en última instancia, no es más que un tipo particular de células cerebrales. Cometen el mismo error simplificador que cometen quienes, tomando como base los escáneres de IRMf, afirman que la actividad en una u otra región explica el «amor» o el «lenguaje». Véase Caramazza *et al.* (2014).

25.

«Ningún hombre es una isla entera por sí mismo; cada hombre es una pieza del continente, una parte del todo». Donne (1839), pp. 574-575. El psicólogo Chris Frith ha llevado esa perspectiva un poco más allá al sostener que la función primaria de toda experiencia consciente es social (Frith, 2007).

26.

Antonio Damasio subraya este aspecto de la yoidad en su libro *La sensación de lo que ocurre* (Damasio, 2000).

27.

James (1890), p. 242.

9. SER UN ANIMAL-MÁQUINA

1.

Anaïs Nin, *La seducción del minotauro* (1961), p. 124. Nin atribuyó la frase a un antiguo texto talmúdico.

2.

El concepto de la Gran Cadena del Ser se originó en la Antigua Grecia con Platón, Aristóteles y Plotino, y alcanzó su máximo desarrollo en la Europa occidental, durante la Edad Media.

3.

Véase un vivaz relato histórico de esta evolución en George Makari, *Alma máquina* (2016).

4.

Véase <**https://en.m.wikipedia.org/wiki/Trademark_argument**>. Véase también Hatfield (2002).

5.

Shugg (1968), p. 279. Véase también la argumentación original en *The Philosophical Works of Descartes,* trad. inglesa de E. S. Haldane y G. R. T. Ross (Nueva York, 1955), vol. 1, pp. 114-116 y 118.

6.

La Mettrie (1748).

7.

Godfrey-Smith (1996); Maturana y Varela (1980).

8.

Craig (2002).

9.

Critchley y Harrison (2013).

10.

Véase más sobre el papel de la ínsula en la interocepción en Barrett y Simmons (2015) y Craig (2009).

11.

James (1884), p. 190. El debate entre los teóricos «clásicos» de las emociones, seguidores de Darwin porque postulan la existencia de unas emociones innatas conservadas por especies sucesivas, y los «constructivistas», que las niegan, continúa aún vivo hoy en día. En el bando de los primeros, tenemos al biólogo Jaak Panksepp y a sus seguidores. Panksepp defendía que hay un conjunto de emociones básicas que se sustancian en unos circuitos neurales específicos y también antiguos desde el punto de vista evolutivo (Panksepp, 2004); véase también Darwin (1872). En el bando de los segundos, se encuentran, por ejemplo, los neurocientíficos Lisa Feldman Barrett y Joe LeDoux, que proponen diferentes versiones de la idea de que las emociones humanas dependen de las evaluaciones cognitivas (una perspectiva que, como veremos, es similar a la mía). Véase más sobre la historia de las teorías de las emociones en Barrett (2016) y LeDoux (2012).

12.

Véase, por ejemplo, Harrison *et al.* (2010).

13.

Schachter y Singer (1962) es la referencia clásica en cuanto a las teorías valorativas.

14.

Dutton y Aron (1974).

15.

Las líneas generales de esta idea aparecieron por vez primera en un artículo de 2011 (Seth *et al.,* 2011b), y se depuraron y concretaron en otro artículo, este de 2013, que, desde entonces, se ha convertido en el de referencia estándar (Seth, 2013). Las ideas básicas se ampliaron a partir de 2015 con la teoría del «animal-máquina» sobre la conciencia y el yo: Seth (2015a); Seth (2019a); Seth y Friston (2016); Seth y Tsakiris (2018).

16.

Las ideas de Damasio tuvieron una honda influencia en mis propias reflexiones sobre este tema, sobre todo en lo relacionado con el yo; véase Damasio (1994, 2000, 2010). Lisa Feldman Barrett pone el énfasis, como yo, en el papel que tienen las predicciones interoceptivas en las emociones: véanse Barrett y Satpute (2019), y Barrett y Simmons (2015), así como su excelente libro *La vida secreta del cerebro. Cómo se construyen las emociones* (Barrett, 2017).

17.

Véase Petzschner *et al.* (2019). Se están empezando a recopilar pruebas adicionales a favor de la inferencia interoceptiva procedentes de estudios con animales. Por ejemplo, dos experimentos recientes parecen sugerir que las neuronas de la ínsula de los ratones codifican algo muy parecido a las predicciones interoceptivas (Gehrlach *et al.,* 2019; Livneh *et al.,* 2020).

18.

Véanse Aspell *et al.* (2013); Suzuki *et al.* (2013). En un estudio relacionado, Micah Allen y un grupo de colaboradores mostraron que la excitación fisiológica inesperada influye en la percepción de estímulos visuales, lo que, una vez más, es indicativo de una interacción entre procesos exteroceptivos e interoceptivos (Allen *et al.,* 2016). Véase un estado de la cuestión en Park y Blanke (2019), y sendos análisis de los problemas de medición de la sensibilidad interoceptiva por medio de la detección de los latidos cardíacos en Brener y Ring (2016), y en Zamariola *et al.* (2018). De la influencia de la sugestibilidad hipnótica en la corporeización ya hemos hablado en el capítulo 8.

19.

Wiener (1948).

20.

La historia de la cibernética no solo muestra que muchas disciplinas académicas que hoy consideramos diferenciadas formaron parte en su día de un mismo enfoque

común, sino que también nos revela que la ciencia puede apartarse en ocasiones de sendas que, de haberse seguido, podrían haber resultado extraordinariamente fecundas en cuanto a generación de conocimientos. Para aprender más sobre esta cuestión, recomiendo el libro de Jean-Pierre Dupuy, *On the Origins of Cognitive Science: The Mechanization of the Mind* (2009).

21.

Conant y Ashby (1970).

22.

Para un análisis más en profundidad de la diferencia entre ser un modelo y tener un modelo, véanse Seth (2015a) y Seth y Tsakiris (2018).

23.

En algunos artículos científicos, me he referido a esa distinción caracterizándola como una diferencia entre una forma *epistémica* (exploratoria, buscadora de información) y otra *instrumental* (orientada a objetivos y al control) de percepción predictiva: Seth (2019a); Seth y Tsakiris (2018); Tschantz *et al.* (2020b).

24.

Ashby (1952). Por ejemplo, la temperatura corporal central humana debe permanecer entre los 32 °C y los 40 °C. Fuera de ese intervalo, sobreviene rápidamente la muerte.

25.

Las acciones externas e internas se distinguen por el tipo de musculatura implicada. Las acciones externas dependen del sistema muscular esquelético (de músculos estriados), mientras que las (intero)acciones internas dependen de los sistemas musculares visceral (de músculos lisos) y cardíaco. Cada uno de esos tipos de músculos son controlados, a su vez, por diferentes ramas del sistema nervioso periférico (el que no incluye ni la médula espinal ni el encéfalo). Los músculos esqueléticos son controlados por el sistema nervioso somático, mientras que los viscerales y los cardíacos son controlados por el sistema nervioso autónomo o vegetativo.

26.

Véase McLeod *et al.* (2003).

27.

Gibson (1979).

28.

Esta teoría se suele resumir con la máxima de que los sistemas de control controlan lo que *perciben,* no lo que *hacen* (Powers, 1973). Véase una formulación más reciente

de la teoría en Marken y Mansell (2013).

29.

No me opongo a la atribución metafórica de formas a las experiencias no visuales. Los dolores pueden ser agudos o romos. Algunos sabores pueden ser punzantes, y también hay emociones que, a veces, se caracterizan de ese modo: los celos, por ejemplo. Pero son experiencias que no tienen formas en el mismo sentido en que las tienen las tazas, los gatos y las mesas de centro cuando los experimentamos visualmente.

30.

Sterling (2012). Véase un modelo informático de control interoceptivo alostático en Tschantz *et al.* (2021).

31.

Véanse versiones más técnicas de esta teoría del animal-máquina y de sus componentes en Seth (2013, 2014b, 2015a y 2019), Seth y Friston (2016), y Seth y Tsakiris (2018). La teoría tiene muchos predecesores e influencias a las que no puedo hacer plenamente justicia aquí. Entre ellos, se incluye el examen filosófico que hizo Thomas Metzinger del yo (Metzinger, 2003a), y las trascendentales descripciones del procesamiento predictivo realizadas por Andy Clark y Jakob Hohwy (Clark, 2016; Hohwy, 2013). La teoría tiene contraída una deuda muy especial con otras proposiciones teóricas que también han postulado la existencia de lazos profundos (aunque diferentes de los descritos aquí) entre la vida, el cuerpo, la mente y la conciencia. En este sentido, me han influido mucho Antonio Damasio (1994 y 2000, p. ej.), Gerald Edelman (1989, p. ej.), Karl Friston (2010, p. ej.), Joe LeDoux (2019, p. ej.) y Evan Thompson (2014, p. ej.; véase también Varela *et al.,* 1993). Véanse ideas relacionadas con estas en Panksepp (2005); Park y Tallon-Baudry (2014); Solms (2021); el concepto de «sesgo de existencia» de Metzinger (2021), y el trabajo de Lisa Feldman Barrett (2017, p. ej.).

32.

La interconexión estrecha entre la conciencia y la regulación fisiológica plantea nuevos interrogantes sobre el papel del tronco del encéfalo (o tallo cerebral: el conjunto de núcleos situados entre las partes más profundas de los hemisferios cerebrales y la médula espinal). Normalmente, se pensaba que el tronco del encéfalo era un «factor posibilitador» de la conciencia en un sentido muy parecido al de cuando se dice que un cable eléctrico es un factor posibilitador para una televisión. Pero el tallo cerebral desempeña un papel muy activo en la regulación fisiológica y eso ha llevado a algunos investigadores a sugerir que es ahí donde nace la conciencia, sin necesidad de que intervenga el córtex (Solms, 2021; véase también Merker, 2007). Creo que eso es harto improbable, dado el peso de las pruebas

explicativas que vinculan la corteza cerebral (y el tálamo) con los estados conscientes. Dicho esto, es muy posible que el tronco del encéfalo tenga un papel más decisivo en la formación de los estados conscientes de lo que cabría deducir a partir de la analogía del cable eléctrico. Véase uno de estos enfoques más matizados en Parvizi y Damasio (2001).

33.

Sería interesante analizar si la ceguera al autocambio se atenúa en épocas de enfermedad o lesión, en las que puede que sea más útil para el cerebro percibir con mayor precisión lo que está ocurriendo en el cuerpo. De hecho, hay un nuevo subcampo de la neurociencia cognitiva que aborda cuestiones de este tipo y que se llama «psicosomática computacional» (Petzschner *et al.,* 2017).

34.

Esto es algo que los griegos ya habían asumido. Aunque se suele asociar a Sócrates con la máxima «conócete a ti mismo», los estoicos pusieron el énfasis en la importancia de la ecuanimidad y el autocontrol. Los defensores de la teoría del control perceptivo podrían ir más lejos aún y afirmar que regulamos nuestra condición fisiológica a fin de percibirnos estables.

35.

Cotard (1880).

36.

Una de las posibles formas en que el yo puede perder su realidad es cuando los modelos generativos subyacentes no codifican (como suelen hacer) el rico repertorio de predicciones interoceptivas condicionales o contrafactuales sobre cómo las diferentes acciones afectan a la regulación fisiológica. Nótese la analogía con la posibilidad de que las predicciones condicionales visuales sean el factor subyacente a la fenomenología de «lo objetual». Véase Seth y Tsakiris (2018).

37.

A la tesis que afirma que la conciencia depende de una propiedad específicamente *biológica* (que un ordenador hecho de silicio jamás podrá tener) se la llama a veces *naturalismo biológico.* He preferido no usar ese término aquí porque diferentes autores lo han empleado con sentidos dispares (véase un comentario sobre ello en Schneider, 2019).

10. PEZ EN EL AGUA

1.

Friston ha publicado un gran número de artículos sobre el principio de la energía

libre. Tiene dos estudios generales clave sobre la cuestión (Friston, 2009 y 2010).

2.

Buckley *et al.* (2017).

3.

Véase <**www.lesswrong.com/posts/wpZJvgQ4HvJE2bysy/god-help-us-let-s-try-to-understand-friston-on-free-energy**>. Otras joyas de este tipo son «How to read Karl Friston (in the original Greek)» («Cómo leer a Karl Friston (en su griego original)»), de Alianna Maren, en <**www.aliannajmaren.com/2017/07/27/how-to-read-karl-friston-in-the-original-greek**>, y «Free Energy: How the f*ck does that work, ecologically» [«Energía libre: Cómo co*o funciona eso, a nivel ecológico»], de Andrew Wilson y Sabrina Golonka, en <**psychsciencenotes.blogspot.com/2016/11/free-energy-how-fck-does-that-work.html**>.

4.

El PEL hace referencia a límites en términos de «mantas de Markov», un concepto tomado de la estadística y del aprendizaje automático. Para un sistema definido por un conjunto de variables aleatorias, una manta de Markov es una partición estadística del sistema en «estados internos», «estados externos» y «estados manta», donde la manta es lo que separa lo interno de lo externo. Las mantas de Markov satisfacen el requisito de que las variables que queden dentro de la manta (los estados internos) sean condicionalmente independientes de las que queden fuera (los estados externos) y viceversa. Esto significa que la dinámica de los estados internos se puede predecir íntegramente a partir de los estados internos y estados manta pasados. Véase más sobre mantas de Markov y el PEL en Kirchhoff *et al.* (2018), y véase también una crítica muy esclarecedora en Bruineberg *et al.* (2020).

5.

En un sentido técnico, la energía libre proporciona un límite superior para una magnitud llamada *sorpresa* o *autoinformación,* que podemos concebir como aquella que especifica lo (estadísticamente) inesperado que es un suceso. Que sea el límite superior significa que la energía libre no puede ser inferior a la sorpresa. La sorpresa está relacionada con la cantidad de entropía teórico-informacional en la que, bajo supuestos de estado estacionario no en equilibrio, la media de sorpresa a largo plazo es la entropía. En un sentido informal, la entropía es como la incertidumbre, y la incertidumbre es la sorpresa media que te esperas encontrar.

6.

Más detalles para los lectores con inclinaciones matemáticas. La energía libre se define en términos de dos distribuciones de probabilidad: (i) una *densidad de*

reconocimiento, que codifica una mejor conjetura actual acerca del estado del entorno, y (ii) una *densidad generativa,* que codifica un modelo probabilístico de cómo los estados ambientales moldean (generan) unos *inputs* sensoriales. Aquí, el concepto de «entorno» alude a las causas ocultas de las señales sensoriales, sean cuales sean. Las cantidades de energía libre tienen dos componentes: una *energía* que se corresponde con la «sorpresa», y una *entropía relativa,* que refleja lo «separada» que la densidad de reconocimiento está de la densidad posterior real (la probabilidad sobre los estados del entorno, dados los *inputs* sensoriales). La distancia entre esas densidades se mide por una magnitud que, en teoría de la información, recibe el nombre de divergencia de Kullback-Leibler (KL). Si suponemos que las densidades de reconocimiento y generativa son gaussianas (y si añadimos algunas suposiciones más relacionadas, por ejemplo, con la independencia de las escalas temporales), entonces la energía libre se corresponde directamente con el error de predicción (ponderado por la precisión) del procesamiento predictivo. Como la menor medida posible de «separación» entre densidades es cero, se entiende que la energía libre siempre es mayor que la sorpresa (es decir, que proporciona un límite superior a esta última). Esto significa, a su vez, que disminuir la energía libre debe comportar la reducción, o bien de la divergencia entre la densidad de reconocimiento y la real posterior (producida por la mejora de las inferencias perceptivas), o bien de la sorpresa (gracias a la toma de muestras de nuevos *inputs* sensoriales).

7.

Jakob Hohwy ha puesto a este proceso el elegante nombre de «autoprobación» (Hohwy, 2014). Desde el punto de vista matemático, esta perspectiva está justificada, ya que minimizar la energía libre equivale a maximizar la evidencia (bayesiana) de los modelos. De hecho, de igual modo que la energía libre proporciona un límite superior a la sorpresa, también nos da un límite inferior para la evidencia de los modelos (el llamado *límite inferior de evidencia* –o ELBO, por sus iniciales en inglés– en el aprendizaje automático); véase Winn y Bishop (2005).

8.

El «problema de la habitación oscura» fue una de las primeras objeciones que se le planteó al PEL (Friston *et al.,* 2012). La misma crítica resurgió mientras escribía este libro, y mis colegas y yo la refutamos de nuevo. Véanse Seth *et al.* (2020); Sun y Fireston (2020).

9.

Hohwy (2020b), p. 9.

10.

Goodstein (1985), p. 1.

11.

Un ejemplo de esta potenciación de la teoría del animal-máquina nos lleva de vuelta a la noción de que los sistemas vivos mantienen un límite entre sí mismos y su entorno, un límite que en el PEL se entiende en términos de una manta de Markov (véase la nota al pie de la página 197). El ejemplo es que, para Friston, la mera existencia o identificación de una manta de Markov implica de forma muy directa que se está produciendo una inferencia activa. Véase Kirchhoff *et al.* (2018), y véase una crítica de nuevo en Bruineberg *et al.* (2020).

12.

Tómense como ejemplo nuestras investigaciones sobre cómo los agentes minimizadores de energía libre aprenden modelos perceptivos adaptativamente sesgados de sus entornos (Tschantz *et al.,* 2020b).

13.

Hohwy y Seth (2020). Ha habido otros intentos de ligar el PEL a la conciencia, por ejemplo en términos de la profundidad temporal de los modelos generativos (Friston, 2018); véanse también Solms (2018); Solms (2021); Williford *et al.* (2018).

14.

Estas tentativas adoptan la forma de una «colaboración entre contrarios», en la que los proponentes de cada teoría estipulan de antemano qué resultado de un experimento respaldará o cuestionará su teoría preferida. Aquí, la colaboración entre antagonistas concreta enfrenta a la TII y a la inferencia activa (más que al PEL en sí). En el capítulo 3, ya mencioné uno de esos experimentos propuestos: la TII predice que la inactivación de neuronas ya inactivas tendrá una incidencia en la percepción consciente, mientras que la inferencia activa hipotetiza que no.

11. GRADOS DE LIBERTAD

1.

McEwan (2000). Vaya por delante mi reconocimiento para Patrick Haggard, del University College de Londres, por señalarme tan acertada cita.

2.

Una certera guía para manejarnos por este campo de minas filosófico es Bayne (2008).

3.

Strawson (2008), p. 367.

4.

Este enfoque, que suscribo, expresa una posición que los filósofos llaman *compatibilismo*. El compatibilismo sostiene que hay ciertas concepciones razonables del libre albedrío que son compatibles con el hecho de que el universo sea determinista. Por el contrario, el libre albedrío *libertario* (no confundir con las filosofías políticas que comparten ese calificativo) es la versión filosófica del libre albedrío espectral. También hay quienes defienden lo que llaman un *determinismo duro,* que sostiene que el determinismo es real, y deduce de ello que ninguna concepción razonable del libre albedrío puede sobrevivir a tal condición. Nótese que yo puedo mostrarme agnóstico al respecto de si el universo es determinista o no (y lo soy), y aun así, ser también un compatibilista.

5.

Aunque el universo sea determinista a cierto nivel fundamental, podría ocurrir que unas fluctuaciones aparentemente aleatorias a nivel de las neuronas y las sinapsis tuvieran un papel significativo en la función cerebral. Es posible, puede que incluso probable, pero, repito, carece de importancia para lo que estoy diciendo aquí.

6.

Los primeros en registrar el potencial de disposición fueron los fisiólogos alemanes Hans Kornhuber y Lüder Deecke en los años sesenta. Le pusieron el nombre de *Bereitschaftspotential* (Kornhuber y Deecke, 1965).

7.

Para los filósofos, las intenciones y los impulsos son cosas diferentes. Puede que yo sienta el *impulso* de golpear a alguien que me está molestando, pero reprimo dicho impulso porque no tengo *intención* de hacer daño a nadie (ni de que me lleven detenido). Las intenciones están sujetas a razones y normas como no lo están los impulsos. En casos simples como el de doblar voluntariamente un dedo, equivalen a más o menos lo mismo, y el propio Libet usó ambos términos indistintamente en la mayoría de casos.

8.

Como era de esperar, tanto el inicio del potencial de disposición como el momento de la intención consciente son previos al movimiento en sí.

9.

Libet *et al.* (1983).

10.

El filósofo Al Mele expuso una parecida duda conceptual unos años antes (Mele, 2009).

11.

Schurger *et al.* (2012).

12.

En Caspar *et al.* (2016), hay un interesante experimento sobre esta cuestión.

13.

Esta cita proviene de un ensayo que Schopenhauer presentó ante la Real Sociedad Noruega de las Ciencias en 1839. Véase una traducción al inglés en Zucker (2013), p. 531. En esa distinción a la que alude Schopenhauer, podemos ver la raíz de todas las adicciones.

14.

Aquí hay, además, un punto de contacto interesante con la cibernética. Ross Ashby, a quien ya conocimos en el capítulo 9 por su «teorema del buen regulador» y su concepto de variable esencial, también es conocido por su «ley de la variedad requerida», que formuló unos años antes (Ashby, 1956). Según esta ley, o principio, para que un sistema de control funcione, debe ser capaz de entrar en, al menos, tantos estados como el entorno que lo está alterando. Según escribió el propio Ashby, «solo la variedad puede doblegar a la variedad». Véase Seth (2015a).

15.

Dennett (1984).

16.

Brass y Haggard (2008); Haggard (2008); Haggard (2019).

17.

Fried *et al.* (1991). Si se estimula con mayor intensidad esa misma área se puede generar tanto un impulso como su acción correspondiente.

18.

Brass y Haggard (2007).

19.

Wegner (2002).

20.

Como dijo Sam Harris en un *podcast* reciente, «el problema no radica solamente en que el problema del libre albedrío no tenga ningún sentido desde el punto de vista objetivo, sino también en que tampoco lo tiene desde el punto de vista subjetivo». Véase <**https://samharris.org/podcasts/241-final-thoughts-on-free-will**>.

21.

Csíkszentmihályi (1990); Harris (2012).

22.

Della Sala *et al.* (1991); Formisano *et al.* (2011).

23.

El caso de la pedofilia inducida por un tumor está muy bien descrito en Burns y Swerdlow (2003). Del de Charles Whitman se ha hablado en múltiples ocasiones, como, por ejemplo, en un artículo de 2011 de la revista *The Atlantic* escrito por David Eagleman: <**www.theatlantic.com/magazine/archive/2011/07/the-brain-on-trial/308520**>.

24.

«No me atribuyo el mérito de nada. Todo está determinado, de principio a fin, por fuerzas sobre las que no tenemos control alguno. Está determinado tanto para el insecto como para la estrella. Los seres humanos, los vegetales o el polvo cósmico bailamos todos al son de una melodía misteriosa». Este es un pasaje en el que vemos expuesta en todo su esplendor la conocida adscripción de Einstein al determinismo. (Otra frase famosa suya fue que «Dios no juega a los dados con el universo», un rechazo de la aleatoriedad inherente a la mecánica cuántica). No obstante, como ya hemos visto, no hace falta suscribir el determinismo para rechazar el libre albedrío espectral. El pasaje procede de una entrevista que le hizo George Sylvester Viereck y que se publicó en el *Saturday Evening Post* el 26 de octubre de 1929 (p. 117).

25.

De la incoherencia de la responsabilidad moral se habla en Waller (2011); en Dennett (1984 y 2003) se expone la visión alternativa. Recientemente, ha aparecido un debate muy esclarecedor sobre esta cuestión (Dennett y Caruso, 2021).

26.

El neurobiólogo Björn Brembs ha argumentado que podemos detectar rastros de eso que llamamos libre albedrío incluso en criaturas muy simples cuando se comportan de maneras variables y (en apariencia) espontáneas que también parecen salirles «de dentro». Puede que, en el fondo, comportamientos de ese tipo (ahí está, por ejemplo, la impredecible «respuesta de huida» de una cucaracha) que son beneficiosos para eludir a los depredadores delaten los orígenes evolutivos de cómo los humanos hemos llegado a controlar nuestros muchos grados de libertad. Véase Brembs (2011).

12. MÁS ALLÁ DE LO HUMANO

1.

Evans (1906).

2.

Descartes publicó sus ideas sobre los animales en el siglo XVII, por lo que coexistieron en parte con esas otras creencias y prácticas medievales. La visión cartesiana se convertiría en dominante posteriormente, a medida que la Ilustración fue arraigando en toda Europa.

3.

Los defensores de las teorías de la conciencia como un «pensamiento de orden superior» (que describí brevemente en el capítulo 1) tal vez discreparían a este respecto. Véase Brown *et al.* (2019).

4.

Desde el extremo opuesto a Descartes, Darwin adoptó una perspectiva marcadamente antropomórfica, sobre todo, en lo referente a la expresión de emociones en los animales (Darwin, 1872). Su presuposición de que existía un conjunto de emociones conservadas a lo largo y ancho de múltiples especies levantó la veda al uso de experimentos con animales para el estudio de las emociones humanas, una línea de trabajo que posteriormente tomarían también Jaak Panksepp y otros investigadores, y que logró fusionarse con la cultura contemporánea, integrada en ella en forma de la asociación de unas determinadas expresiones faciales con unas emociones «innatas» o «básicas», como se puede apreciar en el trabajo de Paul Ekman (1992). Como ya se explicó en el capítulo 9, esta perspectiva de las «emociones básicas» ha sido puesta en cuestión por «constructivistas» contemporáneos como Lisa Feldman Barrett y Joe LeDoux, quienes, como yo mismo, ponen el acento en el papel de las interpretaciones descendentes (generadas «desde arriba») en la formación del contenido consciente. Véanse Barrett (2016); LeDoux (2012).

5.

Véase Steiner y Redish (2014), p. 1001.

6.

El biólogo Ernst Haeckel acuñó el término «biopsiquismo» en 1892 para describir la idea según la cual todos los seres vivos son sintientes y solo ellos lo son, a diferencia de lo que se propugna desde la perspectiva pampsiquista, que viene a afirmar que la conciencia es una propiedad de todas las formas de la materia. Véase Thompson (2007).

7.

Un indicador más sofisticado que el tamaño absoluto del cerebro es el *cociente de encefalización,* que mide el tamaño relativo del cerebro en función del tamaño del cuerpo. Hay mucho debate en torno a si el cociente de encefalización es un predictor fiable de la capacidad cognitiva de las diferentes especies (Herculano-Houzel, 2016; Reep *et al.,* 2007). Lo que no hay, en ningún caso, es motivo suficiente para tomar el

tamaño cerebral o el cociente de encefalización como marcador de la presencia de la conciencia en las diversas especies.

8.

Seth *et al.* (2005).

9.

Kelz y Mashour (2019).

10.

Lyamin *et al.* (2018); Walker (2017).

11.

Uexküll (1957).

12.

Gallup (1970).

13.

Como puede leerse en un artículo que tiene uno de mis títulos favoritos de siempre: «Another gorilla *(Gorilla gorilla gorilla)* recognizes himself in a mirror» (Posada y Colell, 2007).

14.

Véase un estudio general del tema en Gallup y Anderson (2020). El debate sobre el lábrido limpiador se explica muy bien en Kohda *et al.* (2019) y en De Waal (2019).

15.

Gallup y Anderson (2018).

16.

Cowey y Stoerig (1995).

17.

Boly *et al.* (2013).

18.

Kessler y Rawlins (2016).

19.

Véase <**www.ted.com/talks/frans_de_waal_do_animals_have_morals**>.

20.

En un estudio se vio que los monos Rhesus pueden superar la prueba del espejo si se les entrena durante semanas para hacerlo (L. Chang *et al.,* 2017). Pero superar la prueba tras mucho entrenamiento es una cosa bien distinta de usar espontáneamente

los espejos para autorreconocerse.

21.

Borrelli *et al.* (2006). Al llevarme este maravilloso, pero también supervoluminoso, libro de vuelta a casa, superé con mucho el peso máximo permitido para mi billete inicial de Ryanair.

22.

En 2015, me invitaron a colaborar en un libro escribiendo un capítulo sobre la «conciencia extraterrestre». Tras pensármelo bastante, decidí escribir sobre algo tan real como los pulpos, en vez de especular sobre alienígenas potenciales (Seth, 2016a). También se ha explorado ese paralelismo entre pulpos y extraterrestres en la ingeniosa película de Denis Villeneuve, *La llegada* (2016).

23.

Godfrey-Smith (2017), p. 10.

24.

Hochner (2012); Shigeno *et al.* (2018).

25.

Carls-Diamante (2017).

26.

Liscovitch-Brauer *et al.* (2017).

27.

Fiorito y Scotto (1992). Véanse también D. B. Edelman y Seth (2009); Mather (2019), y, como texto clásico sobre el comportamiento de los cefalópodos, Hanlon y Messenger (1996).

28.

Véase <**www.bbc.co.uk/programmes/p05nzfn1**>. Hay unas imágenes no menos impresionantes en el documental de 2020 *My Octopus Teacher,* de Pippa Ehrlich y James Reed, en el que podemos ver cómo el cineasta Craig Foster entabla una relación sorprendentemente estrecha con un pulpo.

29.

El biólogo marino Roger Hanlon ha captado en vídeo muchos ejemplos de camuflaje de los pulpos. El siguiente es uno de los mejores: <**www.youtube.com/watch?v=JSq8nghQZqA**>.

30.

Mather (2019); Messenger (2001).

31.

Van Giesen *et al.* (2020).

32.

Nesher *et al.* (2014).

33.

Véase más sobre la conciencia en animales distintos de los mamíferos en D. B. Edelman y Seth (2009), y D. B. Edelman *et al.* (2005).

34.

Clayton *et al.* (2007); Jao Keehn *et al.* (2019); Pepperberg y Gordon (2005); Pepperberg y Shive (2001).

35.

Curiosamente, si bien los cerebros aviares tienen algo parecido al córtex que poseen los de los mamíferos (lo que, en las aves, se conoce con el nombre de «palio»), carecen de equivalente para el cuerpo calloso, que es lo que conecta los dos hemisferios corticales en las especies de nuestro grupo. Puede que las aves representen una especie de caso natural de «cerebro dividido», lo que plantea interrogantes acerca de la unidad de la conciencia aviar (Xiao y Gunturkun, 2009). El libro *The Thing with Feathers,* de Noah Strycker (2014), es una preciosa introducción a la cognición y la conducta aviares.

36.

Véanse sendos excelentes tratamientos de la evolución de la conciencia en Feinberg y Mallatt (2017); Ginsburg y Jablonka (2019), y LeDoux (2019).

37.

Véase en Birch (2017) un muy útil estudio general del tema, motivado por el argumento de que deberíamos conceder a los animales el «beneficio de la duda» cuando no existen pruebas concluyentes (un argumento que, a nivel formal, recibe el nombre de «principio de precaución»). Es de destacar que la Unión Europea decidió en 2010 incluir los cefalópodos bajo el paraguas protector de la legislación sobre el bienestar animal (directiva 2010/63/EU).

38.

Entler *et al.* (2016).

39.

Khuong *et al.* (2019).

40.

Kelz y Mashour (2019).

41.

Jonathan Birch, Alexandra Schnell y Nicola Clayton proponen el útil concepto de «perfil de conciencia» para caracterizar la variación de la experiencia consciente entre especies. Postulan cinco dimensiones de variación: la riqueza perceptiva, la riqueza evaluativa, la unidad, la temporalidad y la yoidad. Véase Birch *et al.* (2020).

13. MENTES-MÁQUINA

1.

En su libro de 1964 *Dios y Golem, S. A.,* el polímata y pionero Norbert Wiener otorgó a los golems un papel central en sus especulaciones sobre los riesgos de la futura IA.

2.

La parábola del maximizador de clips ideada por Nick Bostrom cuenta la historia de una IA diseñada para fabricar el máximo de clips sujetapapeles posible. En su empeño por cumplir ese cometido, y como carece de valores humanos pero, al mismo tiempo, es muy inteligente, esta IA termina destruyendo el mundo. Véase Bostrom (2014).

3.

Véase una aproximación gratamente serena a la hipótesis de la «singularidad» en Shanahan (2015).

4.

Puede que nos sintamos tentados a afirmar que la conciencia y la inteligencia son doblemente disociables, es decir, que cada una de ellas puede existir sin la otra. Pero eso no sería del todo correcto. Si bien estoy convencido de que la inteligencia puede existir aun en ausencia de conciencia, es posible que la conciencia requiera de un nivel de inteligencia superior a cero.

5.

La idea de una conciencia (y una inteligencia) multidimensional evoca el concepto de «perfil de conciencia» postulado por Jonathan Birch y sus colaboradores (Birch *et al.,* 2020), así como la aproximación multidimensional a los niveles de conciencia en los seres humanos propuesta por Tim Bayne, Jakob Hohwy y Adrian Owen (Bayne *et al.,* 2016).

6.

Dehaene *et al.* (2017). La disponibilidad global se corresponde con la popular teoría del «espacio de trabajo global» de la conciencia, mientras que la «automonitorización» recoge aspectos de las teorías del pensamiento de orden superior. Comentamos brevemente esas dos teorías en el capítulo 1. Los autores del

artículo de *Science* admitían de forma explícita que tal vez estuvieran dejando de lado el componente «experiencial» de la conciencia. Para mí, eso es dejarse mucho en el tintero.

7.

Existe esta posibilidad porque la TII acepta una parte del funcionalismo (la independencia respecto al sustrato) pero no la otra (la suficiencia de que se dé el modo «correcto» de relaciones entre *inputs* y *outputs)*. Algunos mecanismos, entre los que destacan las redes neuronales artificiales de proalimentación que posean un tamaño adecuado, pueden implementar relaciones arbitrariamente complejas entre *inputs* y *outputs*. Estos mecanismos, implementados del modo correcto, pueden tener una apariencia externa de inteligencia y/o conciencia. Pero las redes que son puramente de proalimentación no generan ninguna información integrada (para esto último, siempre se necesita que haya cierta recurrencia o «bucle»). Por consiguiente, desde el punto de vista de la TII, los «zombis conductuales» (que, como expliqué en el capítulo 1, son aquellos artefactos que parecen conscientes desde fuera, pero carecen de toda conciencia por dentro) son perfectamente posibles. Véase Tononi y Koch (2015).

8.

Estas ideas guardan una estrecha relación con el concepto de autopoiesis («autocreación» en griego), desarrollado por el biólogo chileno Humberto Maturana. Un sistema autopoiético es aquel que es capaz de mantenerse y reproducirse por sí mismo, lo que significa que produce también los componentes físicos necesarios para la continuidad de su existencia como sistema. Aunque la autopoiesis es, ante todo, una teoría de la célula, existen ciertos vínculos interesantes entre la autopoiesis celular y el principio de la energía libre (véase el capítulo 10). Ambos sugieren una continuidad estrecha entre la «vida» y la «mente», que, a su vez, da a entender que la mente (y también la conciencia) es algo más que lo que un sistema «hace» (Kirchhoff, 2018; Maturana y Varela, 1980). Tuve la fortuna de conocer a Maturana –que falleció en mayo de 2021, a los noventa y dos años– en enero de 2019, en su Santiago de Chile natal, donde pasamos un rato bebiendo café y hablando sobre ideas como esas en la sombreada terraza de una cafetería del barrio de Providencia.

9.

En el «juego de imitación» original de Alan Turing, hay dos humanos del mismo género y una máquina. La máquina y uno de los humanos –el colaborador– fingen ambos ser una persona del género opuesto. El otro humano tiene que decidir cuál de ellas es la máquina y cuál el colaborador (Turing, 1950).

10.

Este término fue acuñado por Murray Shanahan, cuyo libro *Embodiment and the Inner*

Life (2010) fue una de las fuentes de inspiración de *Ex_Machina*.

11.

Véase <**www.reading.ac.uk/news-archive/press-releases/pr583836.html**>.

12.

«Eugene Goostman is a real boy – the Turing Test says so», *The Guardian,* Pass notes, 9 de junio de 2014. Véase <**https://www.theguardian.com/technology/shortcuts/2014/jun/09/eugene-goostman-turing-test-computer-program**>.

13.

Esa caracterización del test de Turing como un examen de la «credulidad humana» figura en un artículo de 2015 de John Markoff, «Software is smart enough for SAT, but still far from intelligent», *The New York Times,* 21 de septiembre de 2015, <**www.nytimes.com/2015/09/21/technology/personaltech/software-is-smart-enough-for-sat-but-still-far-from-intelligent.html**>.

14.

GPT son las siglas de «Generative Pre-trained Transformer» («Transformador Generativo Preentrenado»), un tipo de red neuronal especializada en predicción y generación lingüísticas. Estas redes son entrenadas mediante un enfoque de aprendizaje profundo no supervisado dirigido esencialmente a que «predigan la palabra siguiente» cuando se les da una palabra o un fragmento de texto previos. GPT-3 cuenta con la friolera de 175 000 millones de parámetros y fue entrenado con unos 45 terabytes de datos de texto. Véase <**https://openai.com/blog/openai-api**>, y véanse detalles técnicos también en <**https://arxiv.org/abs/2005.14165**>.

15.

Obviamente, eso depende de lo que se entienda por «comprender». Habrá quienes digan que la «comprensión» humana no difiere cualitativamente del tipo de «comprensión» manifestado por GPT-3. El científico cognitivo Gary Marcus no está de acuerdo con esa postura y yo coincido con él. Véase <**www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion**>.

16.

«A robot wrote this entire article. Are you scared yet, human?», *The Guardian,* Opinion, 8 de septiembre de 2020. Véase <**www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3**>. No queda claro hasta qué punto este ejemplo es representativo o no.

17.

Becker-Asano *et al.* (2010).

18.

Mori *et al.* (2012).

19.

Hacer un *deepfake* («ultrafalso») es generar un vídeo realista, que no real, de un rostro humano (en la mayoría de casos) usando el aprendizaje automático para combinar un vídeo de origen y otro de destino. Un ejemplo muy difundido en 2017 fueron unos vídeos muy creíbles de Barack Obama diciendo cosas que nunca había dicho (https://www.youtube.com/watch?v=cQ54GDm1eL0). En otra serie de vídeos, estos de Tom Cruise «ultrafalseado» en TikTok en 2021, vemos que el listón se ha subido ya de manera muy sustancial (https://www.theverge.com/22303756/tiktok-tom-cruise-impersonator-deepfake).

20.

El investigador en IA Stuart Russell ha descrito con elocuencia las amenazas planteadas por la IA del momento actual y del futuro inmediato –así como de qué modo se podrían rediseñar los sistemas de IA para eludirlas– en su libro *Human Compatible* (2019). Nina Schick ha hecho un trabajo excelente parecido a propósito del riesgo que encierran los *deepfakes* (Schick, 2020).

21.

«Philosopher Daniel Dennett on AI, robots and religion», *The Financial Times,* 3 de marzo de 2017. Véase <**https://www.ft.com/content/96187a7a-fce5-11e6-96f8-3700c5664d30**>.

22.

Véase Metzinger (2021).

23.

Emmanuelle Charpentier y Jennifer Doudna fueron galardonadas con el Premio Nobel de química de 2020 por sus aportaciones al desarrollo de la técnica CRISPR. La *E. coli* sintética se creó en el laboratorio de Jason Chin; véase Fredens *et al.* (2019).

24.

Trujillo *et al.* (2019).

25.

He investigado la posibilidad de conciencia en organoides en un artículo reciente coescrito con Tim Bayne y Marcello Massimini (Bayne *et al.,* 2020).

26.

Son temas que se están tomando muy en serio. En el verano de 2020, fui invitado (junto con varios neurocientíficos más) a hablar ante el comité conjunto de la Academia Nacional de Estados Unidos, convocado con el objetivo de ayudar a fijar unos marcos regulatorios y legales para las investigaciones tanto sobre organoides como sobre «quimeras» (animales modificados genéticamente para expresar características humanas específicas). Véase <**www.nationalacademies.org/our-work/ethical-legal-and-regulatory-issues-associated-with-neural-chimeras-and-organoids**>.

27.

Carl Zimmer, «Organoids are not brains. How are they making brain waves?», *The New York Times,* 29 de agosto de 2019. Véase <**www.nytimes.com/2019/08/29/science/organoids-brain-alysson-muotri.html**>.

28.

Véase un análisis comedido de las posibilidades y los obstáculos que se plantean para la carga y descarga informática de mentes en Schneider (2019).

29.

La hipótesis de la simulación se presenta con el siguiente argumento (Bostrom, 2003). Una civilización que sea lo bastante futura y haya sabido eludir el peligro de la autoextinción tendría probablemente acceso a unos inmensos recursos informáticos. Algunos miembros de semejante civilización tal vez sentirían curiosidad por ejecutar simulaciones informáticas detalladas de sus antepasados. Dado que se podrían ejecutar muchas de esas simulaciones desde ese distante futuro, para cualquier vida experiencial individual actual sería racional concluir que es mucho más probable que ella misma sea una de esas mentes simuladas que no que sea uno de los humanos biológicos originales. Bostrom lo escribió así: «Si no creemos que estamos viviendo actualmente dentro de una simulación informática, tampoco tenemos derecho a creer que tendremos descendientes que ejecutarán un sinfín de tales simulaciones de sus antepasados» (Bostrom, 2003, p. 243). Uno de los múltiples problemas que tengo con este argumento es que presupone que el funcionalismo es verdad: es decir, que, en lo que a la conciencia se refiere, simulación equivale a sustanciación. Pero, como ya he mencionado aquí, yo pienso que el funcionalismo dista mucho de ser un supuesto sólido y seguro.

EPÍLOGO

1.

Si está conectado aún al riego sanguíneo y continúa estando «vivo», ¿no podría este hemisferio desconectado desarrollar su propia conciencia aislada? Tampoco sería

descartable que surgieran «islas de consciencia» potenciales como esta en otras neurotecnologías incipientes, como los cerebros de cerdo reanimados *ex cranio,* o los organoides cerebrales que he descrito en el capítulo anterior. Comentamos todos estos casos en Bayne *et al.* (2020).

2.

Chalmers (1995b), p. 201.

3.

Véase también Hoffman (2019).

REFERENCIAS BIBLIOGRÁFICAS

Albright, T. D. (2012), «On the perception of probable things: Neural substrates of associative memory, imagery, and perception», *Neuron,* 74, 2, pp. 227-245.

Allen, M., Frank, D., Schwarzkopf, D. S., *et al.* (2016), «Unexpected arousal modulates the influence of sensory noise on confidence», *Elife,* 5, p. e18103.

Anscombe, G. E. M. (1959), *An Introduction to Wittgenstein's Tractatus,* St. Augustine's Press, Londres. [Hay edición en español: *Introducción al* Tractatus *de Wittgenstein,* El Ateneo, Buenos Aires, 1977]

Aru, J., Bachmann, T., Singer, W., *et al.* (2012), «Distilling the neural correlates of consciousness», *Neuroscience and Biobehavioral Reviews,* 36, 2, pp. 737-746.

Ashby, W. R. (1952), *Design for a Brain,* Chapman and Hall, Londres. [Hay edición en español: *Proyecto para un cerebro,* Tecnos, Madrid, 1965]

Ashby, W. R. (1956), *An Introduction to Cybernetics,* Chapman and Hall, Londres. [Hay edición en español: *Introducción a la cibernética,* Nueva Visión, Buenos Aires, 1972]

Aspell, J. E., Heydrich, L., Marillier, G., *et al.* (2013), «Turning the body and self inside out: Visualized heartbeats alter bodily self-consciousness and tactile perception», *Psychological Science,* 24, 12, pp. 2445-2453.

Baars, B. J. (1988), *A Cognitive Theory of Consciousness,* Cambridge University Press, Nueva York.

Barber, T. X. (1961), «Physiological effects of "hipnosis"», *Psychological Bulletin,* 58, pp. 390-419.

Barnes, J. (2008), *Nothing to Be Frightened of,* Knopf, Nueva York. [Hay edición en español: *Nada que temer,* Anagrama, Barcelona, 2010]

Barnett, L., Muthukumaraswamy, S. D., Carhart-Harris, R. L., *et al.* (2020), «Decreased directed functional connectivity in the psychedelic state», *Neuroimage,* 209, p. 116462.

Barrett, A. B., y Seth, A. K. (2011), «Practical measures of integrated information for time-series data», *PLoS Computational Biology,* 7, 1, p. e1001052.

Barrett, L. F. (2017), *How Emotions Are Made: The Secret Life of the Brain,* Houghton Mifflin Harcourt, Boston (Massachusetts). [Hay edición en español: *La vida secreta del cerebro. Cómo se construyen las emociones,* Paidós, Barcelona, 2018]

Barrett, L. F., y Satpute, A. B. (2019), «Historical pitfalls and new directions in the neuroscience of emotion», *Neuroscience Letters,* 693, pp. 9-18.

Barrett, L. F., y Simmons, W. K. (2015), «Interoceptive predictions in the brain», *Nature Reviews Neuroscience,* 16, 7, pp. 419-429.

Bauby, J.-M. (1997), *Le Scaphandre et le Papillon,* Robert Laffont, París. [Hay edición en español: *La escafandra y la mariposa,* Plaza & Janés, Barcelona, 1997]

Bayne, T. (2008), «The phenomenology of agency», *Philosophy Compass,* 3, 1, pp. 182-202.

Bayne, T. (2010), *The Unity of Consciousness,* Oxford University Press, Oxford.

Bayne, T. (2018), «On the axiomatic foundations of the integrated information theory of consciousness», *Neuroscience of Consciousness,* 1, p. niy007.

Bayne, T., Hohwy, J., y Owen, A. M. (2016), «Are there levels of consciousness?», *Trends in Cognitive Sciences,* 20, 6, pp. 405-413.

Bayne, T., Seth, A. K., y Massimini, M. (2020), «Are there islands of awareness?», *Trends in Neurosciences,* 43, 1, pp. 6-16.

Bechtel, W., y Williamson, R. C. (1998), «Vitalism», en E. Craig (ed.),

Routledge Encyclopedia of Philosophy, Routledge, Londres.

Becker-Asano, C., Ogawa, K., Nishio, S., *et al.* (2010), «Exploring the uncanny valley with Geminoid HI-1 in a real-world application», en *IADIS International Conferences Interfaces and Human Computer Interaction,* pp. 121-128.

Berger, J. (1972), *Ways of Seeing,* Penguin, Londres. [Hay edición en español: *Modos de ver,* Gustavo Gili, Barcelona, 1974]

Birch, J. (2017), «Animal sentience and the precautionary principle», *Animal Sentience,* 16, 1.

Birch, J., Schnell, A. K., y Clayton, N. S. (2020), «Dimensions of animal consciousness», *Trends in Cognitive Sciences,* 24, 10, pp. 789-801.

Blake, R., Brascamp, J., y Heeger, D. J. (2014), «Can binocular rivalry reveal neural correlates of consciousness?», *Philosophical Transactions of the Royal Society B: Biological Sciences,* 369, 1641, p. 20130211.

Blanke, O., Landis, T., Spinelli, L., *et al.* (2004), «Out-of-body experience and autoscopy of neurological origin», *Brain,* 127 (2.ª parte), pp. 243-258.

Blanke, O., Slater, M., y Serino, A. (2015), «Behavioral, neural, and computational principles of bodily self-consciousness», *Neuron,* 88, 1, pp. 145-166.

Block, N. (2005), «Two neural correlates of consciousness», *Trends in Cognitive Sciences,* 9, 2, pp. 46-52.

Boly, M., Seth, A. K., Wilke, M., *et al.* (2013), «Consciousness in humans and non-human animals: Recent advances and future directions», *Frontiers in Psychology,* 4, p. 625.

Borrelli, L., Gherardi, F., y Fiorito, G. (2006), *A Catalogue of Body Patterning in Cephalopoda,* Firenze University Press, Florencia.

Bostrom, N. (2003), «Are you living in a computer simulation?», *Philosophical Quarterly,* 53, 11, pp. 243-255.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies,* Oxford University Press, Oxford. [Hay edición en español: *Superinteligencia: Caminos, peligros, estrategias,* Teell, Zaragoza, 2016]

Botvinick, M., y Cohen, J. (1998), «Rubber hands “feel” touch that eyes see», *Nature,* 391, 6669, p. 756.

Brainard, D. H., y Hurlbert, A. C. (2015), «Colour vision: Understanding #TheDress», *Current Biology,* 25, 13, pp. R551-R554.

Brass, M., y Haggard, P. (2007), «To do or not to do: The neural signature of self-control», *Journal of Neuroscience,* 27, 34, pp. 9141-9145.

Brass, M., y Haggard, P. (2008), «The what, when, whether model of intentional action», *Neuroscientist,* 14, 4, pp. 319-325.

Braun, N., Debener, S., Spychala, N., *et al.* (2018), «The senses of agency and ownership: A review», *Frontiers in Psychology,* 9, p. 535.

Brembs, B. (2011), «Towards a scientific concept of free will as a biological trait: Spontaneous actions and decision-making in invertebrates», *Proceedings of the Royal Society B: Biological Sciences,* 278, 1707, pp. 930-939.

Brembs, B. (2020), «The brain as a dynamically active organ», *Biochemical and Biophysical Research Communications,* doi:10.1016/ j.bbrc.2020.12.011.

Brener, J., y Ring, C. (2016), «Towards a psychophysics of interoceptive processes: The measurement of heartbeat detection», *Philosophical Transactions of the Royal Society B: Biological Sciences,* 371, 1708, p. 20160015.

Brown, H., Adams, R. A., Parees, I., *et al.* (2013), «Active inference, sensory attenuation and illusions», *Cognitive Processing,* 14, 4, pp. 411-427.

Brown, R., Lau, H., y LeDoux, J. E. (2019), «Understanding the higher-order approach to consciousness», *Trends in Cognitive Sciences,* 23, 9, pp. 754-768.

Brugger, P., y Lenggenhager, B. (2014), «The bodily self and its disorders: Neurological, psychological and social aspects», *Current Opinion in Neurology,* 27, 6, pp. 644-652.

Bruineberg, J., Dolega, K., Dewhurst, J., *et al.* (2020), «The Emperor’s

new Markov blankets», http://philsci-archive.pitt.edu/18467.

Bruner, J. S., y Goodman, C. C. (1947), «Value and need as organizing factors in perception», *Journal of Abnormal and Social Psychology,* 42, 1, pp. 33-44.

Buckley, C., Kim, C.-S., McGregor, S., y Seth, A. K. (2017), «The free energy principle for action and perception: A mathematical review», *Journal of Mathematical Psychology,* 81, pp. 55-79.

Burns, J. M., y Swerdlow, R. H. (2003), «Right orbitofrontal tumor with pedophilia symptom and constructional apraxia sign», *Archives of Neurology,* 60, 3, pp. 437-440.

Buzsáki, G. (2019), *The Brain from Inside Out,* Oxford University Press, Oxford.

Byrne, A., y Hilbert, D. (2011), «Are colors secondary qualities?», en L. Nolan (ed.), *Primary and Secondary Qualities: The Historical and Ongoing Debate,* Oxford University Press, Oxford, pp. 339-361.

Caramazza, A., Anzellotti, S., Strnad, L., *et al.* (2014), «Embodied cognition and mirror neurons: A critical assessment», *Annual Review of Neuroscience,* 37, pp. 1-15.

Carhart-Harris, R. L., Erritzoe, D., Williams, T., *et al.* (2012), «Neural correlates of the psychedelic state as determined by fMRI studies with psilocybin», *Proceedings of the National Academy of Sciences of the USA,* 109, 6, pp. 2138-2143.

Carls-Diamante, S. (2017), «The octopus and the unity of consciousness», *Biology and Philosophy,* 32, pp. 1269-1287.

Casali, A. G., Gosseries, O., Rosanova, M., *et al.* (2013), «A theoretically based index of consciousness independent of sensory processing and behavior», *Science Translational Medicine,* 5, 198, p. 198ra105.

Casarotto, S., Comanducci, A., Rosanova, M., *et al.* (2016), «Stratification of unresponsive patients by an independently validated index of brain complexity», *Annals of Neurology,* 80, 5, pp. 718-729.

Caspar, E. A., Christensen, J. F., Cleeremans, A., *et al.* (2016), «Coercion changes the sense of agency in the human brain», *Current Biology,* 26,

5, pp. 585-592.
Chalmers, D. J. (1995a), «Facing up to the problem of consciousness», *Journal of Consciousness Studies,* 2, 3, pp. 200-219.
Chalmers, D. J. (1995b), «The puzzle of conscious experience», *Scientific American,* 273, 6, pp. 80-86.
Chalmers, D. J. (1996), *The Conscious Mind: In Search of a Fundamental Theory,* Oxford University Press, Nueva York. [Hay edición en español: *La mente consciente. En busca de una teoría fundamental,* Gedisa, Barcelona, 1999]
Chalmers, D. J. (2018), «The meta-problem of consciousness», *Journal of Consciousness Studies,* 25, 9-10, pp. 6-61.
Chang, H. (2004), *Inventing Temperature: Measurement and Scientific Progress,* Oxford University Press, Nueva York.
Chang, L., Zhang, S., Poo, M. M., *et al.* (2017), «Spontaneous expression of mirror self-recognition in monkeys after learning precise visual-proprioceptive association for mirror images», *Proceedings of the National Academy of Sciences of the USA,* 114, 12, pp. 3258-3263.
Churchland, P. S. (1996), «The hornswoggle problem», *Journal of Consciousness Studies,* 3, 5-6, pp. 402-408.
Cisek, P. (2007), «Cortical mechanisms of action selection: The affordance competition hypothesis», *Philosophical Transactions of the Royal Society B: Biological Sciences,* 362, 1485, pp. 1585-1599.
Clark, A. (2013), «Whatever next? Predictive brains, situated agents, and the future of cognitive science», *Behavioral and Brain Sciences,* 36, 3, pp. 181-204.
Clark, A. (2016), *Surfing Uncertainty,* Oxford University Press, Oxford.
Clayton, N. S., Dally, J. M., y Emery, N. J. (2007), «Social cognition by food-caching corvids: The western scrub-jay as a natural psychologist», *Philosophical Transactions of the Royal Society B: Biological Sciences,* 362, 1480, pp. 507-522.
Cobb, M. (2020), *The Idea of the Brain: A History,* Profile Books, Londres.
Collier, R. (2012), «Hospital-induced delirium hits hard», *Canadian*

Medical Association Journal, 184, 1, pp. 23-24.

Conant, R., y Ashby, W. R. (1970), «Every good regulator of a system must be a model of that system», *International Journal of Systems Science,* 1, 2, pp. 89-97.

Cotard, J. (1880), «Du délire hypocondriaque dans une forme grave de la mélancolie anxieuse. Mémoire lu à la Société médico-psychologique dans la séance du 28 Juin 1880», *Annales Medico-Psychologiques,* pp. 168-174.

Cowey, A., y Stoerig, P. (1995), «Blindsight in monkeys», *Nature,* 373, 6511, pp. 247-249.

Craig, A. D. (2002), «How do you feel? Interoception: The sense of the physiological condition of the body», *Nature Reviews Neuroscience,* 3, 8, pp. 655-666.

Craig, A. D. (2009), «How do you feel—now? The anterior insula and human awareness», *Nature Reviews Neuroscience,* 10, 1, pp. 59-70.

Craver, C., y Tabery, J. (2017), «Mechanisms in science», en *The Stanford Ecyclopedia of Philosophy,* plato.stanford.edu/entries/science-mechanisms.

Crick, F., y Koch, C. (1990), «Towards a neurobiological theory of consciousness», *Seminars in the Neurosciences,* 2, pp. 263-275.

Critchley, H. D., y Harrison, N. A. (2013), «Visceral influences on brain and behavior», *Neuron,* 77, 4, pp. 624-638.

Csikszentmihalyi, M. (1990), *Flow: The Psychology of Optimal Experience,* Harper & Row, Nueva York. [Hay edición en español: *Fluir* (Flow). *Una psicología de la felicidad,* Kairós, Barcelona, 1997]

Damasio, A. (1994), *Descartes' Error,* Macmillan, Londres. [Hay edición en español: *El error de Descartes. La emoción, la razón y el cerebro,* Crítica, Barcelona, 1996]

Damasio, A. (2000), *The Feeling of What Happens: Body and Emotion in the Making of Consciousness,* Harvest Books. [Hay edición en español: *La sensación de lo que ocurre. Cuerpo y emoción en la construcción de la conciencia,* Debate, Madrid, 2001]

Damasio, A. (2010), *Self Comes to Mind: Constructing the Conscious Brain,* William Heinemann, Londres. [Hay edición en español: *Y el cerebro creó al hombre. ¿Cómo pudo el cerebro generar emociones, sentimientos, ideas y el yo?,* Destino, Barcelona, 2010]

Darwin, C. (1872), *The Expression of Emotions in Man and Animals,* Fontana Press, Londres. [Hay edición en español: *La expresión de las emociones en los animales y en el hombre,* Alianza, Madrid, 1984]

Davis, D. H., Muniz-Terrera, G., Keage, H. A., *et al.* (2017), «Association of delirium with cognitive decline in late life: A neuropathologic study of three population-based cohort studies», *JAMA Psychiatry,* 74, 3, pp. 244-251.

De Graaf, T. A., Hsieh, P. J., y Sack, A. T. (2012), «The "correlates" in neural correlates of consciousness», *Neuroscience and Biobehavioral Reviews,* 36, 1, pp. 191-197.

De Haan, E. H., Pinto, Y., Corballis, P. M., *et al.* (2020), «Split-brain: What we know about cutting the corpus callosum now and why this is important for understanding consciousness», *Neuropsychological Review,* 30, pp. 224-233.

De Lange, F. P., Heilbron, M., y Kok, P. (2018), «How do expectations shape perception?», *Trends in Cognitive Sciences,* 22, 9, pp. 764-779.

De Waal, F. B. M. (2019), «Fish, mirrors, and a gradualist perspective on self-awareness», *PLoS Biology,* 17, 2, p. e3000112.

Dehaene, S., y Changeux, J. P. (2011), «Experimental and theoretical approaches to conscious processing», *Neuron,* 70, 2, pp. 200-227.

Dehaene, S., Lau, H., y Kouider, S. (2017), «What is consciousness, and could machines have it?», *Science,* 358, 6362, pp. 486-492.

Deisseroth, K. (2015), «Optogenetics: Ten years of microbial opsins in neuroscience», *Nature Neuroscience,* 18, 9, pp. 1213-1225.

Della Sala, S., Marchetti, C., y Spinnler, H. (1991), «Right-sided anarchic (alien) hand: A longitudinal study», *Neuropsychologia,* 29, 11, pp. 1113-1127.

Demertzi, A., Tagliazucchi, E., Dehaene, S., *et al.* (2019), «Human

consciousness is supported by dynamic complex patterns of brain signal coordination», *Science Advances,* 5, 2, p. eaat7603.

Dennett, D. C. (1984), *Elbow Room: The Varieties of Free Will Worth Wanting,* MIT Press, Cambridge (Massachusetts). [Hay edición en español: *La libertad de acción. Un análisis de la exigencia de libre albedrío,* Gedisa, Barcelona, 1992]

Dennett, D. C. (1991), *Consciousness Explained,* Little, Brown, Boston (Massachusetts). [Hay edición en español: *La conciencia explicada. Una teoría interdisciplinar,* Paidós, Barcelona, 1995]

Dennett, D. C. (1998), «The myth of double transduction», en S. Hameroff, A. W. Kasniak, y A. C. Scott (eds.), *Toward a Science of Consciousness II: The Second Tucson Discussions and Debates,* MIT Press, Cambridge (Massachusetts), pp. 97-101.

Dennett, D. C. (2003), *Freedom Evolves,* Penguin Books, Nueva York. [Hay edición en español: *La evolución de la libertad,* Paidós, Barcelona, 2004]

Dennett, D. C. (2015), «Why and how does consciousness seem the way it seems?», en T. Metzinger y J. M. Windt (eds.), *Open MIND,* MIND Group, Fráncfort.

Dennett, D. C. y Caruso, G. (2021), *Just Deserts: Debating Free Will,* Polity, Cambridge.

Deutsch, D. (2012), *The Beginning of Infinity. Explanations that Transform the World,* Penguin Books, Nueva York. [Hay edición en español: *El comienzo del infinito. Explicaciones que transforman el mundo,* Biblioteca Buridán, Barcelona, 2012]

DiNuzzo, M., y Nedergaard, M. (2017), «Brain energetics during the sleep-wake cycle», *Current Opinion in Neurobiology,* 47, pp. 65-72.

Donne, J. (1839), «Devotions upon emergent occasions: Meditation XVII» [1624], en H. Alford (ed.), *The Works of John Donne,* Henry Parker, Londres, vol. 3, pp. 574-575. [Hay edición en español de las *Meditaciones* de John Donne: *Meditaciones en tiempos de crisis,* Ariel, Barcelona, 2012]

Duffy, S. W., Vulkan, D., Cuckle, H., *et al.* (2020), «Effect of mammographic screening from age forty years on breast cancer mortality (UK Age trial): Final results of a randomised, controlled trial», *Lancet Oncology,* 21, 9, pp. 1165-1172.

Dupuy, J.-P. (2009), *On the Origins of Cognitive Science: The Mechanization of the Mind,* 2.ª ed., MIT Press, Cambridge (Massachusetts).

Dutton, D. G., y Aron, A. P. (1974), «Some evidence for heightened sexual attraction under conditions of high anxiety», *Journal of Personal and Social Psychology,* 30, 4, pp. 510-517.

Edelman, D. B., Baars, B. J., y Seth, A. K. (2005), «Identifying hallmarks of consciousness in non-mammalian species», *Consciousness and Cognition,* 14, 1, pp. 169-187.

Edelman, D. B., y Seth, A. K. (2009), «Animal consciousness: A synthetic approach», *Trends in Neuroscience,* 32, 9, pp. 476-484.

Edelman, G. M. (1989), *The Remembered Present,* Basic Books, Nueva York.

Edelman, G. M., y Gally, J. (2001), «Degeneracy and complexity in biological systems», *Proceedings of the National Academy of Sciences of the USA,* 98, 24, pp. 13763-13768.

Ehrsson, H. H. (2007), «The experimental induction of out-of-body experiences», *Science,* 317, 5841, p. 1048.

Ekman, P. (1992), «An argument for basic emotions», *Cognition and Emotion,* 6, 3-4, pp. 169-200.

Entler, B. V., Cannon, J. T., y Seid, M. A. (2016), «Morphine addiction in ants: A new model for self-administration and neurochemical analysis», *Journal of Experimental Biology,* 219 (18.ª parte), pp. 2865-2869.

Evans, E. P. (1906), *The Criminal Prosecution and Capital Punishment of Animals,* William Heinemann, Londres.

Feinberg, T. E., y Mallatt, J. M. (2017), *The Ancient Origins of Consciousness: How the Brain Created Experience,* MIT Press, Cambridge

(Massachusetts).

Feldman, H., y Friston, K. J. (2010), «Attention, uncertainty, and free-energy», *Frontiers in Human Neuroscience,* 4, p. 215.

Felleman, D. J., y Van Essen, D. C. (1991), «Distributed hierarchical processing in the primate cerebral cortex», *Cerebral Cortex,* 1, 1, pp. 1-47.

Ferrarelli, F., Massimini, M., Sarasso, S., *et al.* (2010), «Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness», *Proceedings of the National Academy of Sciences of the USA,* 107, 6, pp. 2681-2686.

Fiorito, G., y Scotto, P. (1992), «Observational learning in *Octopus vulgaris», Science,* 256, 5056, pp. 545-547.

Firestone, C. (2013), «On the origin and status of the "El Greco fallacy"», *Perception,* 42, 6, pp. 672-674.

Fletcher, P. C., y Frith, C. D. (2009), «Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia», *Nature Reviews Neuroscience,* 10, 1, pp. 48-58.

Fleming, S. M. (2020), «Awareness as inference in a higher-order state space», *Neuroscience of Consciousness,* 2020, 1, p. niz020.

Flounders, M. W., Gonzalez-Garcia, C., Hardstone, R., *et al.* (2019), «Neural dynamics of visual ambiguity resolution by perceptual prior», *Elife,* 8, p. e41861.

Formisano, R., D'Ippolito, M., Risetti, M., *et al.* (2011), «Vegetative state, minimally conscious state, akinetic mutism and Parkinsonism as a continuum of recovery from disorders of consciousness: An exploratory and preliminary study», *Functional Neurology,* 26, 1, pp. 15-24.

Frankish, K. (2017), *Illusionism as a Theory of Consciousness,* Imprint Academic, Exeter.

Frässle, S., Sommer, J., Jansen, A., *et al.* (2014), «Binocular rivalry: Frontal activity relates to introspection and action but not to perception», *Journal of Neuroscience,* 34, 5, pp. 1738-1747.

Fredens, J., Wang, K., De la Torre, D., *et al.* (2019), «Total synthesis of *Escherichia coli* with a recoded genome», *Nature,* 569, 7757, pp. 514-518.

Fried, I., Katz, A., McCarthy, G., *et al.* (1991), «Functional organization of human supplementary motor cortex studied by electrical stimulation», *Journal of Neuroscience,* 11, 11, pp. 3656-3666.

Friston, K. J. (2009), «The free-energy principle: A rough guide to the brain?», *Trends in Cognitive Sciences,* 13, 7, pp. 293-301.

Friston, K. J. (2010), «The free-energy principle: A unified brain theory?», *Nature Reviews Neuroscience,* 11, 2, pp. 127-138.

Friston, K. J. (2018), «Am I self-conscious? (Or does self-organization entail self-consciousness?)», *Frontiers in Psychology,* 9, p. 579.

Friston, K. J., Daunizeau, J., Kilner, J., *et al.* (2010), «Action and behavior: A free-energy formulation», *Biological Cybernetics,* 102, 3, pp. 227-260.

Friston, K. J., Thornton, C., y Clark, A. (2012), «Free-energy minimization and the dark-room problem», *Frontiers in Psychology,* 3, p. 130.

Frith, C. D. (2007), *Making Up the Mind: How the Brain Creates Our Mental World,* Wiley-Blackwell, Oxford.

Gallagher, S. (2008), «Direct perception in the intersubjective context», *Consciousness and Cognition,* 17, 2, pp. 535-543.

Gallese, V., Fadiga, L., Fogassi, L., *et al.* (1996), «Action recognition in the premotor cortex», *Brain,* 119 (2.ª parte), pp. 593-609.

Gallup, G. G. (1970), «Chimpanzees: self-recognition», *Science,* 167, pp. 86-87.

Gallup, G. G., y Anderson, J. R. (2018), «The "olfactory mirror" and other recent attempts to demonstrate self-recognition in non-primate species», *Behavioural Processes,* 148, pp. 16-19.

Gallup, G. G., y Anderson, J. R. (2020), «Self-recognition in animals: Where do we stand fifty years later? Lessons from cleaner wrasse and other species», *Psychology of Consciousness: Theory, Research, and*

Practice, 7, 1, pp. 46-58.

Gasquet, J. (1991), *Cézanne: A Memoir with Conversations,* Thames & Hudson Ltd., Londres. [Hay edición en español: *Cézanne. Lo que vi y lo que me dijo,* Gadir, Madrid, 2005]

Gehrlach, D. A., Dolensek, N., Klein, A. S., *et al.* (2019), «Aversive state processing in the posterior insular cortex», *Nature Neuroscience,* 22, 9, pp. 1424-1437.

Gibson, J. J. (1979), *The Ecological Approach to Visual Perception,* Lawrence Erlbaum, Hillsdale (Nueva Jersey).

Gidon, A., Zolnik, T. A., Fidzinski, P., *et al.* (2020), «Dendritic action potentials and computation in human layer 2/3 cortical neurons», *Science,* 367, 6473, pp. 83-87.

Gifford, C., y Seth, A. K. (2013), *Eye Benders: The Science of Seeing and Believing,* Thames & Hudson, Londres.

Ginsburg, S., y Jablonka, E. (2019), *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness,* MIT Press, Cambridge (Massachusetts).

Godfrey-Smith, P. G. (1996), «Spencer and Dewey on life and mind», en M. Boden (ed.), *The Philosophy of Artificial Life,* Oxford University Press, Oxford, pp. 314-331. [Hay edición en español: *Filosofía de la inteligencia artificial,* Fondo de Cultura Económica, Ciudad de México, 1990]

Godfrey-Smith, P. G. (2017), *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness,* Farrar, Strauss, and Giroux, Nueva York.

Goff, P. (2019), *Galileo's Error: Foundations for a New Science of Consciousness,* Rider, Londres.

Gombrich, E. H. (1961), *Art and Illusion: A Study in the Psychology of Pictorial Representation,* Princeton University Press, Ewing (Nueva Jersey). [Hay edición en español: *Arte e ilusión: Estudio sobre la psicología de la representación pictórica,* Gustavo Gili, Barcelona, 1979]

Goodstein, D. L. (1985), *States of Matter,* Courier Corporation, Chelmsford (Massachusetts).

Gregory, R. L. (1980), «Perceptions as hypotheses», *Philosophical Transactions of the Royal Society B*: *Biological Sciences,* 290, 1038, pp. 181-97.

Grill-Spector, K., y Malach, R. (2004), «The human visual cortex», *Annual Review of Neuroscience,* 27, pp. 649-677.

Haggard, P. (2008), «Human volition: Towards a neuroscience of will», *Nature Reviews Neuroscience,* 9, 12, 934-946.

Haggard, P. (2019), «The neurocognitive bases of human volition», *Annual Review of Psychology,* 70, pp. 9-28.

Hanlon, J., y Messenger, J. B. (1996), *Cephalopod Behaviour,* Cambridge University Press, Cambridge.

Harding, D. E. (1961), *On Having No Head,* The Shollond Trust, Londres. [Hay edición en español: *Vivir sin cabeza. Una experiencia zen,* Kairós, Barcelona, 1994]

Harris, S. (2012), *Free Will,* Deckle Edge, Nueva York.

Harrison, N. A., Gray, M. A., Gianaros, P. J., *et al.* (2010), «The embodiment of emotional feelings in the brain», *Journal of Neuroscience,* 30, 38, pp. 12878-12884.

Harvey, I. (2008), «Misrepresentations», en S. Bullock, J. Noble, R. Watson, *et al.* (eds.), *Artificial Life XI: Proceedings of the 11th International Conference On the Simulation and Synthesis of Living Systems,* MIT Press, Cambridge (Massachusetts), pp. 227-233.

Hatfield, G. (2002), *Descartes and the Meditations,* Routledge, Abingdon.

Haun, A. M. (2021), «What is visible across the visual field?», *Neuroscience of Consciousness,* 1, p. niab006.

Haun, A. M., y Tononi, G. (2019), «Why does space feel the way it does? Towards a principled account of spatial experience», *Entropy,* 21, 12, p. 1160.

He, K., Zhang, X., Ren, S., *et al.* (2016), «Deep residual learning for image recognition», en *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 770-778.

Heilbron, M., Richter, D., Ekman, M., *et al.* (2020), «Word contexts

enhance the neural representation of individual letters in early visual cortex», *Nature Communications,* 11, 1, p. 321.

Herculano-Houzel, S. (2009), «The human brain in numbers: A linearly scaled-up primate brain», *Frontiers in Human Neuroscience,* 3, p. 31.

Herculano-Houzel, S. (2016), *The Human Advantage: A New Understanding of How Our Brain Became Remarkable,* MIT Press, Cambridge (Massachusetts). [Hay edición en español: *La ventaja humana. Una nueva interpretación del carácter extraordinario del cerebro humano,* Buridán, Barcelona, 2018]

Hochner, B. (2012), «An embodied view of octopus neurobiology», *Current Biology,* 22, 20, pp. R887-R892.

Hoel, E. P., Albantakis, L., y Tononi, G. (2013), «Quantifying causal emergence shows that macro can beat micro», *Proceedings of the National Academy of Sciences of the USA,* 110, 49, pp. 19790-19795.

Hoffman, D. (2019), *The Case against Reality: Why Evolution Hid the Truth from Our Eyes,* W. W. Norton & Company, Londres.

Hoffman, D., Singh, M., y Prakash, C. (2015), «The interface theory of perception», *Psychonomic Bulletin and Review,* 22, pp. 1480-1506.

Hohwy, J. (2013), *The Predictive Mind,* Oxford University Press, Oxford.

Hohwy, J. (2014), «The self-evidencing brain», *Nous,* 50, 2, pp. 259-285.

Hohwy, J. (2020a), «New directions in predictive processing», *Mind and Language,* 35, 2, pp. 209-223.

Hohwy, J. (2020b), «Self-supervision, normativity and the free energy principle», *Synthese,* doi:10.1007/s11229-020-02622-2.

Hohwy, J., y Seth, A. K. (2020), «Predictive processing as a systematic basis for identifying the neural correlates of consciousness», *Philosophy and the Mind Sciences,* 1, 2, p. 3.

Hurley, S., y Noë, A. (2003), «Neural plasticity and consciousness», *Biology and Philosophy,* 18, pp. 131-168.

Husserl, E. (1960 [1931]), *Cartesian Meditations: An Introduction to Phenomenology,* Nijhoff, La Haya. [Hay edición en español: *Meditaciones cartesianas: Introducción a la fenomenología,* Fondo de

Cultura Económica, Ciudad de México, 1985]

Inagaki, K., y Hatano, G. (2004), «Vitalistic causality in young children's naive biology», *Trends in Cognitive Sciences,* 8, 8, pp. 356-362.

James, W. (1884), «What is an emotion?», *Mind,* 9, 34, pp. 188-205.

James, W. (1890), *The Principles of Psychology,* Henry Holt, Nueva York. [Hay edición en español: *Principios de psicología,* Fondo de Cultura Económica, Ciudad de México, 1983]

Jao Keehn, R. J., Iversen, J. R., Schulz, I., *et al.* (2019), «Spontaneity and diversity of movement to music are not uniquely human», *Current Biology,* 29, 13, pp. R621-R622.

Jensen, F. V. (2000), *Introduction to Bayesian Networks,* Springer, Nueva York.

Jensen, M. P., Jamieson, G. A., Lutz, A., *et al.* (2017), «New directions in hypnosis research: Strategies for advancing the cognitive and clinical neuroscience of hypnosis», *Neuroscience of Consciousness,* 3, 1, p. nix004.

Kail, P. J. E. (2007), *Projection and Realism in Hume's Philosophy,* Oxford University Press, Oxford.

Kandel, E. R. (2012), *The Age of Insight: The Quest to Understand the Unconscious in Art, Mind, and Brain, from Vienna 1900 to the Present.* New York: Random House. [Hay edición en español: *La era del inconsciente: La exploración del inconsciente en el arte, la mente y el cerebro,* Paidós, Barcelona, 2013]

Kelz, M. B., y Mashour, G. A. (2019), «The biology of general anesthesia from paramecium to primate», *Current Biology,* 29, 22, pp. R1199-R1210.

Kessler, M. J., y Rawlins, R. G. (2016), «A seventy-five-year pictorial history of the Cayo Santiago rhesus monkey colony», *American Journal of Primatology,* 78, 1, pp. 6-43.

Khuong, T. M., Wang, Q. P., Manion, J., *et al.* (2019), «Nerve injury drives a heightened state of vigilance and neuropathic sensitization in *Drosophila», Science Advances,* 5, 7, p. eaaw4099.

Kirchhoff, M. (2018), «Autopoeisis, free-energy, and the life-mind continuity thesis», *Synthese,* 195, 6, pp. 2519-2540.

Kirchhoff, M., Parr, T., Palacios, E., *et al.* (2018), «The Markov blankets of life: Autonomy, active inference and the free energy principle», *Journal of the Royal Society Interface,* 15, 138, p. 20170792.

Koch, C. (2019), *The Feeling of Life Itself: Why Consciousness Is Widespread But Can't Be Computed,* MIT Press, Cambridge (Massachusetts).

Kohda, M., Hotta, T., Takeyama, T., *et al.* (2019), «If a fish can pass the mark test, what are the implications for consciousness and self-awareness testing in animals?», *PLoS Biology,* 17, 2, p. e3000021.

Kornhuber, H. H., y Deecke, L. (1965), «Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale», *Pflügers Archiv für die gesamte Physiologie des Menschen und der Tiere,* 284, pp. 1-17.

Konkoly, K. R., Appel, K., Chabani, E., *et al.* (2021), «Real-time dialogue between experimenters and dreamers during REM sleep», *Current Biology,* 31, 7, pp. 1417-1427.

Kuhn, G., Amlani, A. A., y Rensink, R. A. (2008), «Towards a science of magic», *Trends in Cognitive Sciences,* 12, 9, pp. 349-354.

Lakatos, I. (1978), *The Methodology of Scientific Research Programmes: Philosophical Papers,* Cambridge University Press, Cambridge. [Hay edición en español: *La metodología de los programas de investigación científica,* Alianza, Madrid, 1983]

La Mettrie, J. O. de (1748), *L'Homme machine,* Luzac, Leiden. [Hay edición en español: *El hombre máquina,* Valdemar, Madrid, 2000]

Lau, H., y Rosenthal, D. (2011), «Empirical support for higher-order theories of conscious awareness», *Trends in Cognitive Sciences,* 15, 8, pp. 365-373.

LeDoux, J. (2012), «Rethinking the emotional brain», *Neuron,* 73, 4, pp. 653-676.

LeDoux, J. (2019), *The Deep History of Ourselves: The Four-billion-year Story of How We Got Conscious Brains,* Viking, Nueva York. [Hay

edición en español: *Una historia natural de la humanidad. El apasionante recorrido de la vida hasta alcanzar nuestro cerebro consciente,* Paidós, Barcelona, 2021]

LeDoux, J., Michel, M., y Lau, H. (2020), «A little history goes a long way toward understanding why we study consciousness the way we do today», *Proceedings of the National Academy of Sciences of the USA,* 117, 13, pp. 6976-6984.

Lemon, R. N., y Edgley, S. A. (2010), «Life without a cerebellum», *Brain,* 133 (3.ª parte), pp. 652-654.

Lenggenhager, B., Tadi, T., Metzinger, T., *et al.* (2007), «Video ergo sum: Manipulating bodily self-consciousness», *Science,* 317, 5841, pp. 1096-1099.

Lettvin, J. Y. (1976), «On seeing sidelong», *The Sciences,* 16, pp. 10-20.

Libet, B. (1985), «Unconscious cerebral initiative and the role of conscious will in voluntary action», *Behavioral and Brain Sciences,* 8, pp. 529-66.

Libet, B., Wright, E. W., Jr., y Gleason, C. A. (1983), «Preparation- or intention-to-act, in relation to pre-event potentials recorded at the vertex», *Electroencephalography and Clinical Neurophysiology,* 56, 4, pp. 367-372.

Lipton, P. (2004), *Inference to the Best Explanation,* Routledge, Abingdon.

Liscovitch-Brauer, N., Alon, S., Porath, H. T., *et al.* (2017), «Trade-off between transcriptome plasticity and genome evolution in cephalopods», *Cell,* 169, 2, pp. 191-202 e111.

Livneh, Y., Sugden, A. U., Madara, J. C., *et al.* (2020), «Estimation of current and future physiological states in insular cortex», *Neuron,* 105, 6, pp. 1094-1111.e10.

Luppi, A. I., Craig, M. M., Pappas, I., *et al.* (2019), «Consciousness-specific dynamic interactions of brain integration and functional diversity», *Nature Communications,* 10, 1, p. 4616.

Lush, P. (2020), «Demand characteristics confound the rubber hand illusion», *Collabra Psychology,* 6, p. 22.

Lush, P., Botan, V., Scott, R. B., *et al.* (2020), «Trait phenomenological control predicts experience of mirror synaesthesia and the rubber hand illusion», *Nature Communications,* 11, 1, p. 4853.

Lyamin, O. I., Kosenko, P. O., Korneva, S. M., *et al.* (2018), «Fur seals suppress REM sleep for very long periods without subsequent rebound», *Current Biology,* 28, 12, pp. 2000-2005e2002.

Makari, G. (2016), *Soul Machine: The Invention of the Modern Mind,* W. W. Norton, Londres. [Hay edición en español: *Alma máquina. La invención de la mente moderna,* Sexto Piso, Madrid y Ciudad de México, 2021]

Marken, R. S., y Mansell, W. (2013), «Perceptual control as a unifying concept in psychology», *Review of General Psychology,* 17, 2, pp. 190-195.

Markov, N. T., Vezoli, J., Chameau, P., *et al.* (2014), «Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex», *Journal of Comparative Neurology,* 522, 1, pp. 225-259.

Marr, D. (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information,* Freeman, Nueva York. [Hay edición en español: *Visión. Una investigación basada en el cálculo acerca de la representación y el procesamiento humano de la información visual,* Alianza, Madrid, 1982]

Mashour, G. A., Roelfsema, P., Changeux, J. P., *et al.* (2020), «Conscious processing and the global neuronal workspace hypothesis», *Neuron,* 105, 5, pp. 776-798.

Massimini, M., Ferrarelli, F., Huber, R., *et al.* (2005), «Breakdown of cortical effective connectivity during sleep», *Science,* 309, 5744, pp. 2228-2232.

Mather, J. (2019), «What is in an octopus's mind?», *Animal Sentience,* 26, 1, pp. 1-29.

Maturana, H., y Varela, F. (1980), *Autopoiesis and Cognition: The Realization of the Living,* D. Reidel, Dordrecht. [Hay edición en español: *De máquinas y seres vivos. Autopoiesis. La organización de lo*

vivo, Lumen y Editorial Universitaria, Buenos Aires y Santiago de Chile, 2003]

McEwan, I. (2000), *Atonement,* Anchor Books, Nueva York. [Hay edición en español: *Expiación,* Barcelona, Anagrama, 2002]

McGinn, C. (1989), «Can we solve the mind-body problem?», *Mind,* 98, 391, pp. 349-366.

McGrayne, S. B. (2012), *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy,* Yale University Press, New Haven (Connecticut). [Hay edición en español: *La teoría que nunca murió. De cómo la regla de Bayes permitió descifrar el código Enigma, perseguir los submarinos rusos y emerger triunfante de dos siglos de controversia,* Crítica, Barcelona, 2012]

McLeod, P., Reed, N., y Dienes, Z. (2003), «Psychophysics: How fielders arrive in time to catch the ball», *Nature,* 426, 6964, pp. 244-245.

Mediano, P. A. M., Seth, A. K., y Barrett, A. B. (2019), «Measuring integrated information: Comparison of candidate measures in theory and simulation», *Entropy,* 21, 1, p. 17.

Mele, A. (2009), *Effective Intentions: The Power of Conscious Will,* Oxford University Press, Nueva York. [Hay edición en español: *Intenciones efectivas: El poder de la voluntad consciente,* Herder, Barcelona, 2018]

Melloni, L., Schwiedrzik, C. M., Muller, N., *et al.* (2011), «Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness», *Journal of Neuroscience,* 31, 4, pp. 1386-1396.

Merker, B. (2007), «Consciousness without a cerebral cortex: A challenge for neuroscience and medicine», *Behavioral and Brain Sciences,* 30, 1, pp. 63-81; debate en pp. 81-134.

Merleau-Ponty, M. (1962), *Phenomenology of Perception,* Routledge & Kegan Paul, Londres. [Hay edición en español: *Fenomenología de la percepción,* Península, Barcelona, 1975]

Merleau-Ponty, M. (1964), «Eye and mind», en J. E. Edie (ed.), *The Primacy of Perception,* Northwestern University Press, Evanston

(Illinois), pp. 159-190.

Messenger, J. B. (2001), «Cephalopod chromatophores: neurobiology and natural history», *Biological Reviews of the Cambridge Philosophical Society,* 76, 4, pp. 473-528.

Metzinger, T. (2003a), *Being No One,* MIT Press, Cambridge (Massachusetts).

Metzinger, T. (2003b), «Phenomenal transparency and cognitive self-reference», *Phenomenology and the Cognitive Sciences,* 2, pp. 353-393.

Metzinger, T. (2021), «Artificial suffering: An argument for a global moratorium on synthetic phenomenology», *Journal of Artificial Intelligence and Consciousness,* 8, 1, pp. 1-24.

Monroe, R. (1971), *Journeys out of the Body,* Anchor Press, Londres. [Hay edición en español: *Viajes fuera del cuerpo. La expansión de la consciencia más allá de la materia,* Palmyra, Madrid, 2008]

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., *et al.* (2010), «Willful modulation of brain activity in disorders of consciousness», *New England Journal of Medicine,* 362, 7, pp. 579-589.

Mori, M., MacDorman, K. F., y Kageki, N. (2012), «The Uncanny Valley», *IEEE Robotics & Automation Magazine,* 19, 2, pp. 98-100.

Myles, P. S., Leslie, K., McNeil, J., *et al.* (2004), «Bispectral index monitoring to prevent awareness during anaesthesia: The B-Aware randomised controlled trial», *The Lancet,* 363, 9423, pp. 1757-1763.

Naci, L., Sinai, L., y Owen, A. M. (2017), «Detecting and interpreting conscious experiences in behaviorally non-responsive patients», *Neuroimage,* 145 (parte B), pp. 304-313.

Nagel, T. (1974), «What is it like to be a bat?», *Philosophical Review,* 83, 4, pp. 435-450.

Nasraway, S. S., Jr., Wu, E. C., Kelleher, R. M., *et al.* (2002), «How reliable is the Bispectral Index in critically ill patients? A prospective, comparative, single-blinded observer study», *Critical Care Medicine,* 30, 7, pp. 1483-1487.

Nesher, N., Levy, G., Grasso, F. W., *et al.* (2014), «Self-recognition

mechanism between skin and suckers prevents octopus arms from interfering with each other», *Current Biology,* 24, 11, pp. 1271-1275.

Nin, A. (1961), *Seduction of the Minotaur,* Swallow Press, Denver (Colorado). [Hay edición en español: *La seducción del minotauro,* Grijalbo, Barcelona, 1981]

O'Regan, J. K. (2011), *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness,* Oxford University Press, Oxford.

O'Regan, J. K., y Noë, A. (2001), «A sensorimotor account of vision and visual consciousness», *Behavioral and Brain Sciences,* 24, 5, pp. 939-973; debate en pp. 973-1031.

Orne, M. T. (1962), «On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications», *American Psychologist,* 17, pp. 776-783.

Owen, A. M. (2017), *Into the Gray Zone: A Neuroscientist Explores the Border between Life and Death,* Faber & Faber, Londres.

Owen, A. M., Coleman, M. R., Boly, M., *et al.* (2006), «Detecting awareness in the vegetative state», *Science,* 313, 5792, p. 1402.

Palmer, C. E., Davare, M., y Kilner, J. M. (2016), «Physiological and perceptual sensory attenuation have different underlying neurophysiological correlates», *Journal of Neuroscience,* 36, 42, pp. 10803-10812.

Palmer, C. J., Seth, A. K., y Hohwy, J. (2015), «The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism», *Consciousness and Cognition,* 36, pp. 376-389.

Panksepp, J. (2004), *Affective Neuroscience: The Foundations of Human and Animal Emotions,* Oxford University Press, Oxford.

Panksepp, J. (2005), «Affective consciousness: Core emotional feelings in animals and humans», *Consciousness and Cognition,* 14, 1, pp. 30-80.

Park, H. D., y Blanke, O. (2019), «Coupling inner and outer body for self-consciousness», *Trends in Cognitive Sciences,* 23, 5, pp. 377-388.

Park, H. D., y Tallon-Baudry, C. (2014), «The neural subjective frame: From bodily signals to perceptual consciousness», *Philosophical*

Transactions of the Royal Society B: *Biological Sciences,* 369, 1641, p. 20130208.

Parvizi, J., y Damasio, A. (2001), «Consciousness and the brainstem», *Cognition,* 79, 1-2, pp. 135-160.

Penrose, R. (1989), *The Emperor's New Mind,* Oxford University Press, Oxford. [Hay edición en español: *La nueva mente del emperador,* Grijalbo Mondadori, Barcelona, 1991]

Pepperberg, I. M., y Gordon, J. D. (2005), «Number comprehension by a grey parrot (*Psittacus erithacus*), including a zero-like concept», *Journal of Comparative Psychology,* 119, 2, pp. 197-209.

Pepperberg, I. M., y Shive, H. R. (2001), «Simultaneous development of vocal and physical object combinations by a grey parrot *(Psittacus erithacus):* Bottle caps, lids, and labels», *Journal of Comparative Psychology,* 115, 4, pp. 376-384.

Petkova, V. I., y Ehrsson, H. H. (2008), «If I were you: Perceptual illusion of body swapping», *PLoS One,* 3, 12, p. e3832.

Petzschner, F. H., Weber, L. A., Wellstein, K. V., *et al.* (2019), «Focus of attention modulates the heartbeat evoked potential», *Neuroimage,* 186, pp. 595-606.

Petzschner, F. H., Weber, L. A. E., Gard, T., *et al.* (2017), «Computational psychosomatics and computational psychiatry: Toward a joint framework for differential diagnosis», *Biological Psychiatry,* 82, 6, pp. 421-430.

Phillips, M. L., Medford, N., Senior, C., *et al.* (2001), «Depersonalization disorder: Thinking without feeling», *Psychiatry Research,* 108, 3, pp. 145-160.

Pinto, Y., van Gaal, S., de Lange, F. P., *et al.* (2015), «Expectations accelerate entry of visual stimuli into awareness», *Journal of Vision*, 15, 8, p. 13.

Pollan, M. (2018), *How to Change Your Mind,* Penguin, Nueva York. [Hay edición en español: *Cómo cambiar tu mente. Lo que la nueva ciencia de la psicodelia nos enseña sobre la conciencia, la muerte, la adicción, la*

depresión y la trascendencia, Debate, Barcelona, 2018]

Portin, P. (2009), «The elusive concept of the gene», *Hereditas,* 146, 3, pp. 112-117.

Posada, S., y Colell, M. (2007), «Another gorilla *(Gorilla gorilla gorilla)* recognizes himself in a mirror», *American Journal of Primatology,* 69, 5, pp. 576-583.

Powers, W. T. (1973), *Behavior: The Control of Perception,* Aldine de Gruyter, Hawthorne (Nueva York).

Press, C., Kok, P., y Yon, D. (2020), «The perceptual prediction paradox», *Trends in Cognitive Sciences,* 24, 1, pp. 13-24.

Pressnitzer, D., Graves, J., Chambers, C., *et al.* (2018), «Auditory perception: Laurel and Yanny together at last», *Current Biology,* 28, 13, pp. R739-R741.

Raccah, O., Block, N., y Fox, K. (2021), «Does the prefrontal cortex play an essential role in consciousness? Insights from intracranial electrical stimulation of the human brain», *Journal of Neuroscience,* 41, 1, pp. 2076-2087.

Rao, R. P., y Ballard, D. H. (1999), «Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects», *Nature Neuroscience,* 2, 1, pp. 79-87.

Reep, R. L., Finlay, B. L., y Darlington, R. B. (2007), «The limbic system in mammalian brain evolution», *Brain, Behavior and Evolution,* 70, 1, pp. 57-70.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., *et al.* (2019), «A deep learning framework for neuroscience», *Nature Neuroscience,* 22, 11, pp. 1761-1770.

Riemer, M., Trojan, J., Beauchamp, M., *et al.* (2019), «The rubber hand universe: On the impact of methodological differences in the rubber hand illusion», *Neuroscience and Biobehavioral Reviews,* 104, pp. 268-280.

Rosas, F., Mediano, P. A. M., Jensen, H. J., *et al.* (2021), «Reconciling emergences: An information-theoretic approach to identify causal

emergence in multivariate data», *PLoS Computational Biology,* 16, 12, p. e1008289.

Roseboom, W., Fountas, Z., Nikiforou, K., *et al.* (2019), «Activity in perceptual classification networks as a basis for human subjective time perception», *Nature Communications,* 10, 1, p. 267.

Rousseau, M. C., Baumstarck, K., Alessandrini, M., *et al.* (2015), «Quality of life in patients with locked-in syndrome: Evolution over a six-year period», *Orphanet Journal of Rare Diseases,* 10, p. 88.

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control,* Viking, Nueva York.

Sabra, A. I. (1989), *The Optics of Ibn Al-Haytham, Books I-III,* The Warburg Institute, Londres.

Schachter, S., y Singer, J. E. (1962), «Cognitive, social, and physiological determinants of emotional state», *Psychological Review,* 69, pp. 379-399.

Schartner, M. M., Carhart-Harris, R. L., Barrett, A. B., *et al.* (2017a), «Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin», *Scientific Reports,* 7, p. 46421.

Schartner, M. M., Pigorini, A., Gibbs, S. A., *et al.* (2017b), «Global and local complexity of intracranial EEG decreases during NREM sleep», *Neuroscience of Consciousness,* 3, 1, p. niw022.

Schartner, M. M., Seth, A. K., Noirhomme, Q., *et al.* (2015), «Complexity of multi-dimensional spontaneous EEG decreases during propofol induced general anaesthesia», *PLoS One,* 10, 8, p. e0133532.

Schick, N. (2020), *Deep Fakes and the Infocalypse: What You Urgently Need to Know,* Monoray, Monterrey (California).

Schneider, S. (2019), *Artificial You: AI and the Future of Your Mind,* Princeton University Press, Princeton (Nueva Jersey). [Hay edición en español: *Inteligencia artificial. Una exploración filosófica sobre el futuro de la mente y la conciencia,* Kōan, Badalona, 2021]

Schurger, A., Sitt, J. D., y Dehaene, S. (2012), «An accumulator model for spontaneous neural activity prior to self-initiated movement»,

Proceedings of the National Academy of Sciences of the USA, 109, 42, pp. E2904-E2913.

Searle, J. (1980), «Minds, brains, and programs», *Behavioral and Brain Sciences,* 3, 3, pp. 417-457.

Seth, A. K. (2009), «Explanatory correlates of consciousness: Theoretical and computational challenges», *Cognitive Computation,* 1(1), 50-63.

Seth, A. K. (2010), «Measuring autonomy and emergence via Granger causality», *Artificial Life,* 16, 2, pp. 179-196.

Seth, A. K. (2013), «Interoceptive inference, emotion, and the embodied self», *Trends in Cognitive Sciences,* 17, 11, pp. 565-573.

Seth, A. K. (2014a), «Darwin's neuroscientist: Gerald M. Edelman, 1929-2014», *Frontiers in Psychology,* 5, p. 896.

Seth, A. K. (2014b), «A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synaesthesia», *Cognitive Neuroscience,* 5, 2, pp. 97-118.

Seth, A. K. (2015a), «The cybernetic bayesian brain: From interoceptive inference to sensorimotor contingencies», en J. M. Windt y T. Metzinger (eds.), *Open MIND,* 35(T), MIND Group, Fráncfort, pp. 1-24.

Seth, A. K. (2015b), «Inference to the best prediction», en T. Metzinger y J. M. Windt (eds.), *Open MIND,* 35(R), MIND Group, Fráncfort, pp. 1-8.

Seth, A. K. (2016a), «Aliens on earth: What octopus minds can tell us about alien consciousness», en J. Al-Khalili (ed.), *Aliens,* Profile Books, Londres, pp. 47-58.

Seth, A. K. (2016b), «The real problem», *Aeon,* aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one.

Seth, A. K. (2017), «The fall and rise of consciousness science», en A. Haag (ed.), *The Return of Consciousness,* Ax:Son Johnson Foundation, Riga, pp. 13-41.

Seth, A. K. (2018), «Consciousness: The last 50 years (and the next)», *Brain and Neuroscience Advances,* 2, p. 2398212818816019.

Seth, A. K. (2019a), «Being a beast machine: The origins of selfhood in

control-oriented interoceptive inference», en M. Colombo, L. Irvine, y M. Stapleton (eds.), *Andy Clark and his Critics,* Wiley-Blackwell, Oxford, pp. 238-254.

Seth, A. K. (2019b), «From unconscious inference to the Beholder's Share: Predictive perception and human experience», *European Review,* 273, 3, pp. 378-410.

Seth, A. K. (2019c), «Our inner universes», *Scientific American,* 321, 3, pp. 40-47.

Seth, A. K., Baars, B. J., y Edelman, D. B. (2005), «Criteria for consciousness in humans and other mammals», *Consciousness and Cognition,* 14, 1, pp. 119-139.

Seth, A. K., Barrett, A. B., y Barnett, L. (2011a), «Causal density and integrated information as measures of conscious level», *Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences,* 369, 1952, pp. 3748-3767.

Seth, A. K., Dienes, Z., Cleeremans, A., *et al.* (2008), «Measuring consciousness: Relating behavioural and neurophysiological approaches», *Trends in Cognitive Sciences,* 12, 8, pp. 314-321.

Seth, A. K., y Friston, K. J. (2016), «Active interoceptive inference and the emotional brain», *Philosophical Transactions of the Royal Society B: Biological Sciences,* 371, 1708, p. 20160007.

Seth, A. K., Izhikevich, E., Reeke, G. N., *et al.* (2006), «Theories and measures of consciousness: An extended framework», *Proceedings of the National Academy of Sciences of the USA,* 103, 28, pp. 10799-10804.

Seth, A. K., Millidge, B., Buckley, C. L., *et al.* (2020), «Curious inferences: Reply to Sun and Firestone on the dark room problem», *Trends in Cognitive Sciences,* 24, 9, pp. 681-683.

Seth, A. K., Roseboom, W., Dienes, Z., y Lush, P. (2021), «What's up with the rubber hand illusion?», https://psyarxiv.com/b4qcy.

Seth, A. K., Suzuki, K., y Critchley, H. D. (2011b), «An interoceptive predictive coding model of conscious presence», *Frontiers in*

Psychology, 2, p. 395.

Seth, A. K., y Tsakiris, M. (2018), «Being a beast machine: the somatic basis of selfhood», *Trends in Cognitive Sciences,* 22, 11, pp. 969-981.

Shanahan, M. P. (2010), *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds,* Oxford University Press, Oxford.

Shanahan, M. P. (2015), *The Technological Singularity,* MIT Press, Cambridge (Massachusetts).

Sherman, M. T., Fountas, Z., Seth, A. K., *et al.* (2020), «Accumulation of salient events in sensory cortex activity predicts subjective time», www.biorxiv.org/content/10.1101/2020.01.09.900423v4.

Shigeno, S., Andrews, P. L. R., Ponte, G., *et al.* (2018), «Cephalopod brains: An overview of current knowledge to facilitate comparison with vertebrates», *Frontiers in Physiology,* 9, p. 952.

Shugg, W. (1968), «The cartesian beast-machine in English literature (1663-1750)», *Journal of the History of Ideas,* 29, 2, pp. 279-292.

Silver, D., Schrittwieser, J., Simonyan, K., *et al.* (2017), «Mastering the game of Go without human knowledge», *Nature,* 550, 7676, pp. 354-359.

Simons, D. J., y Chabris, C. F. (1999), «Gorillas in our midst: Sustained inattentional blindness for dynamic events», *Perception,* 28, 9, pp. 1059-1074.

Solms, M. (2018), «The hard problem of consciousness and the free energy principle», *Frontiers in Physiology,* 9, p. 2714.

Solms, M. (2021), *The Hidden Spring: A Journey to the Source of Consciousness,* Profile Books, Londres.

Stein, B. E., y Meredith, M. A. (1993), *The Merging of the Senses,* MIT Press, Cambridge (Massachusetts).

Steiner, A. P., y Redish, A. D. (2014), «Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task», *Nature Neuroscience,* 17, 7, pp. 995-1002.

Sterling, P. (2012), «Allostasis: A model of predictive regulation»,

Physiology and Behavior, 106, 1, pp. 5-15.

Stetson, C., Fiesta, M. P., y Eagleman, D. M. (2007), «Does time really slow down during a frightening event?», *PLoS One,* 2, 12, p. e1295.

Stoelb, B. L., Molton, I. R., Jensen, M. P., *et al.* (2009), «The efficacy of hypnotic analgesia in adults: A review of the literature», *Contemporary Hypnosis,* 26, 1, pp. 24-39.

Stoljar, D. (2017), «Physicalism», en E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy,* invierno de 2017, plato.stanford.edu/archives/win2017/entries/physicalism.

Strawson, G. (2008), *Real Materialism and Other Essays,* Oxford University Press, Oxford.

Strycker, N. (2014), *The Thing with Feathers: The Surprising Lives of Birds and What They Reveal about Being Human,* Riverhead Books, Nueva York.

Suárez-Pinilla, M., Nikiforou, K., Fountas, Z., *et al.* (2019), «Perceptual content, not physiological signals, determines perceived duration when viewing dynamic, natural scenes», *Collabra Psychology,* 5, 1, p. 55.

Sun, Z., y Firestone, C. (2020), «The dark room problem», *Trends in Cognitive Sciences,* 24, 5, pp. 346-348.

Sutherland, S. (1989), *International Dictionary of Psychology,* Crossroad Classic, Nueva York.

Suzuki, K., Garfinkel, S. N., Critchley, H. D., y Seth, A. K. (2013), «Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion», *Neuropsychologia,* 51, 13, pp. 2909-2917.

Suzuki, K., Roseboom, W., Schwartzman, D. J., y Seth, A. K. (2017), «A deep-dream virtual reality platform for studying altered perceptual phenomenology», *Scientific Reports,* 7, 1, p. 15982.

Suzuki, K., Schwartzman, D. J., Augusto, R., y Seth, A. K. (2019), «Sensorimotor contingency modulates breakthrough of virtual 3D objects during a breaking continuous flash suppression paradigm»,

Cognition, 187, pp. 95-107.

Suzuki, K., Wakisaka, S., y Fujii, N. (2012), «Substitutional reality system: A novel experimental platform for experiencing alternative reality», *Scientific Reports,* 2, p. 459.

Swanson, L. R. (2016), «The predictive processing paradigm has roots in Kant», *Frontiers in Systems Neuroscience,* 10, p. 79.

Teasdale, G. M., y Murray, L. (2000), «Revisiting the Glasgow Coma Scale and Coma Score», *Intensive Care Medicine,* 26, 2, pp. 153-154.

Teufel, C., y Fletcher, P. C. (2020), «Forms of prediction in the nervous system», *Nature Reviews Neuroscience,* 21, 4, pp. 231-242.

Thompson, E. (2007), *Mind in Life: Biology, Phenomenology, and the Sciences of Mind,* Harvard University Press, Cambridge (Massachusetts).

Thompson, E. (2014), *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy,* Columbia University Press, Nueva York.

Timmermann, C., Roseman, L., Schartner, M., *et al.* (2019), «Neural correlates of the DMT experience assessed with multivariate EEG», *Scientific Reports,* 9, 1, p. 16324.

Tong, F. (2003), «Out-of-body experiences: From Penfield to present», *Trends in Cognitive Sciences,* 7, 3, pp. 104-106.

Tononi, G. (2008), «Consciousness as integrated information: A provisional manifesto», *Biological Bulletin,* 215, 3, pp. 216-242.

Tononi, G. (2012), «Integrated information theory of consciousness: An updated account», *Archives italiennes de biologie,* 150, 4, pp. 293-329.

Tononi, G., Boly, M., Massimini, M., *et al.* (2016), «Integrated information theory: From consciousness to its physical substrate», *Nature Reviews Neuroscience,* 17, 7, pp. 450-461.

Tononi, G., y Edelman, G. M. (1998), «Consciousness and complexity», *Science,* 282, 5395, pp. 1846-1851.

Tononi, G., y Koch, C. (2015), «Consciousness: Here, there and everywhere?», *Philosophical Transactions of the Royal Society B*:

Biological Sciences, 370, 1668.

Tononi, G., Sporns, O., y Edelman, G. M. (1994), «A measure for brain complexity: Relating functional segregation and integration in the nervous system», *Proceedings of the National Academy of Sciences of the USA,* 91, 11, pp. 5033-5037.

Trujillo, C. A., Gao, R., Negraes, P. D., *et al.* (2019), «Complex oscillatory waves emerging from cortical organoids model early human brain network development», *Cell Stem Cell,* 25, 4, pp. 558-569e557.

Tschantz, A., Barca, L., Maisto, D., *et al.* (2021), «Simulating homeostatic, allostatic and goal-directed forms of interoceptive control using active inference», https://www.biorxiv.org/content/10.1101/2021.02.16.431365v1.

Tschantz, A., Millidge, B., Seth, A. K., *et al.* (2020a), «Reinforcement learning through active inference», doi:https://arxiv.org/abs/2002.12636.

Tschantz, A., Seth, A. K., y Buckley, C. (2020b), «Learning action-oriented models», *PLoS Computational Biology,* 16, 4, p. e1007805.

Tsuchiya, N., Wilke, M., Frässle, S., *et al.* (2015), «No-report paradigms: Extracting the true neural correlates of consciousness», *Trends in Cognitive Sciences,* 19, 12, pp. 757-770.

Tulving, E. (1985), «Memory and consciousness», *Canadian Psychology,* 26, pp. 1-12.

Turing, A. M. (1950), «Computing machinery and intelligence», *Mind,* 59, pp. 433-460.

Uexküll, J. v. (1957), «A stroll through the worlds of animals and men: A picture book of invisible worlds», en C. Schiller (ed.), *Instinctive Behavior: The Development of a Modern Concept,* International Universities Press, Nueva York, p. 5.

Van Giesen, L., Kilian, P. B., Allard, C. A. H., *et al.* (2020), «Molecular basis of chemotactile sensation in octopus», *Cell,* 183, 3, pp. 594-604e514.

Van Rijn, H., Gu, B. M., y Meck, W. H. (2014), «Dedicated clock/timing-circuit theories of time perception and timed performance», *Advances in Experimental Medicine and Biology,* 829, pp. 75-99.

Varela, F. J. (1996), «Neurophenomenology: A methodological remedy for the hard problem», *Journal of Consciousness Studies,* 3, pp. 330-350.

Varela, F., Thompson, E., y Rosch, E. (1993), *The Embodied Mind: Cognitive Science and Human Experience,* MIT Press, Cambridge (Massachusetts). [Hay edición en español: *De cuerpo presente. Las ciencias cognitivas y la experiencia humana,* Gedisa, Barcelona, 1997]

Walker, M. (2017), *Why We Sleep,* Scribner, Nueva York. [Hay edición en español: *Por qué dormimos. La nueva ciencia del sueño,* Capitán Swing, Madrid, 2019]

Waller, B. (2011), *Against Moral Responsibility,* MIT Press, Cambridge (Massachusetts).

Wearing, D. (2005), *Forever Today: A Memoir of Love and Amnesia,* Corgi, Londres.

Wegner, D. (2002), *The Illusion of Conscious Will,* MIT Press, Cambridge (Massachusetts).

Weiser, T. G., Regenbogen, S. E., Thompson, K. D., *et al.* (2008), «An estimation of the global volume of surgery: A modelling strategy based on available data», *The Lancet,* 372, 9633, pp. 139-144.

Wheeler, J. A. (1989), «Information, physics, quantum: The search for links», en *Proceedings III International Symposium on Foundations of Quantum Mechanics,* Tokio, pp. 354-358.

Wiener, N. (1948), *Cybernetics: Or Control and Communication in the Animal and Machine,* MIT Press, Cambridge (Massachusetts). [Hay edición en español: *Cibernética: O el control y comunicación en animales y máquinas,* Tusquets, Barcelona, 1985]

Wiener, N. (1964), *God and Golem, Inc.,* MIT Press, Cambridge (Massachusetts). [Hay edición en español: *Dios y Golem, S. A.,* Siglo XXI, Ciudad de México, 1967]

Williford, K., Bennequin, D., Friston, K., *et al.* (2018), «The projective

consciousness model and phenomenal selfhood», *Frontiers in Psychology,* 9, p. 2571.

Winn, J., y Bishop, C. M. (2005), «Variational message passing», *Journal of Machine Learning Research,* 6, pp. 661-694.

Wittmann, M. (2013), «The inner sense of time: How the brain creates a representation of duration», *Nature Reviews Neuroscience,* 14, 3, pp. 217-223.

Witzel, C., Racey, C., y O'Regan, J. K. (2017), «The most reasonable explanation of "the dress": Implicit assumptions about illumination», *Journal of Vision,* 17, 2, p. 1.

Xiao, Q., y Gunturkun, O. (2009), «Natural split-brain? Lateralized memory for task contingencies in pigeons», *Neuroscience Letters,* 458, 2, pp. 75-78.

Zamariola, G., Maurage, P., Luminet, O., *et al.* (2018), «Interoceptive accuracy scores from the heartbeat counting task are problematic: Evidence from simple bivariate correlations», *Biological Psychology,* 137, pp. 12-17.

Zucker, M. (1945), *The Philosophy of American History,* vol. 1: *The Historical Field Theory,* Arnold-Howard, Nueva York.

NOTAS AL PIE

*

Este artículo es uno de los más influyentes de toda la filosofía de la mente. Según Nagel, «un organismo tiene estados mentales conscientes si y solo si hay algo que es cómo es *ser* ese organismo, algo que es cómo es ser *para* ese organismo». Nagel (1974), p. 2 (la cursiva es la del original). [Todas las notas al pie son del autor, a menos que se indique lo contrario]

*

Cada hemisferio de la corteza cerebral contiene cuatro lóbulos. Los lóbulos frontales se encuentran delante. Los parietales se sitúan en la zona posterior, aunque más hacia los lados. Los occipitales están en la parte posterior, más centrados, y los temporales, a los lados, cerca de las orejas. Hay quienes hablan de la existencia de un quinto lóbulo –el límbico– en una zona más profunda del cerebro.

*

Un helicóptero, que sí puede volar marcha atrás, no es un avión. Curiosamente, me alegré de descubrir que el origen de la palabra «helicóptero» no es un combinación de «heli» y de «cóptero», como siempre había supuesto, sino de *helico* («espiral») y *ptero* («ala»). Ahora el nombre tiene mucho más sentido para mí.

*

Se estima que el cerebro humano adulto contiene 86 000 millones de neuronas, y unas mil veces más conexiones. Si las contases a un ritmo de una conexión cada segundo, tardarías casi tres millones de años en acabar. Además, cada vez es más evidente que incluso las neuronas son capaces de llevar a cabo individualmente funciones altamente complejas.

*

El córtex visual está en el lóbulo occipital, en la parte posterior del cerebro.

*

La imagen por resonancia magnética funcional (IRMf) mide una señal metabólica (la oxigenación de la sangre) relativa a la actividad neural: aporta un elevado nivel de detalle espacial, pero solo guarda una relación indirecta con lo que hacen las neuronas. La EEG mide las minúsculas señales eléctricas generadas por la actividad de grandes poblaciones de neuronas próximas a la superficie cortical. Esta técnica registra la actividad cerebral de manera más directa que la IRMf, pero con una menor especificidad espacial.

*

El imaginar (y ejecutar) movimientos fluidos activa unas regiones corticales, como el área motora suplementaria, mientras que el desplazarse por espacios enciende otras regiones, como el giro parahipocampal. Anatómicamente hablando, estas áreas cerebrales están bastante distantes entre sí. De todos modos, ambos ejercicios de imaginación activan regiones que están implicadas en el sentido del oído y en el procesamiento lingüístico, lo cual no debería sorprendernos.

*

La comprensión reflexiva se conserva en el raro estado de los «sueños lúcidos», en el que la persona que sueña es consciente de lo está haciendo y puede orientar voluntariamente su conducta. En un estudio reciente muy interesante, unos investigadores lograron comunicarse con unas personas mientras estas estaban teniendo sueños lúcidos: para ello, aprovecharon sus movimientos oculares como canal comunicativo, de un modo muy similar a como se hizo con los pacientes con enclaustramiento antes mencionados.[24] Estos soñadores y soñadoras fueron capaces de dar respuestas correctas a problemas matemáticos simples y a diversas preguntas contestables con un sí o un no.

*

En su álbum de debut de 1967, *Gorilla,* la Bonzo Dog Doo-Dah Band parodió la música *jazz* tradicional tratando de tocarla lo peor posible. Los gorilas volverán a aparecer más adelante en este libro, por cierto.

*

En teoría de la información, el «bit» es la unidad informacional básica.

*

La idea –tan familiar como completamente equivocada– de que los seres humanos tenemos solamente cinco sentidos se remonta al tratado aristotélico *Sobre el alma,* escrito en torno al 350 a. C.

*

El origen de esta expresión se remonta, como mínimo, a un seminario impartido en la década de 1990 por Ramesh Jain. He intentado rastrear algún posible origen anterior sin éxito.

*

Se puede ver una imagen en color de *El vestido* en <**https://en.wikipedia.org/wiki/The_dress**>. ¿Tú qué ves?

*

Cuando el conocimiento no afecta a la percepción, decimos que esta es «cognitivamente impenetrable».

*

Figuras 6 y 7, tomadas de Teufel, C., Dakin S. C., y Fletcher, P. C. (2018), «Prior object-knowledge sharpens properties of early visual feature detectors», *Scientific Reports*, 8, p. 10 853. Utilizadas con permiso y de los autores y una Licencia Internacional de Atribución 4.0 Creative Commons, y con mi más sincero agradecimiento a Christoph Teufel.

*

A veces, se usa el término *varianza* en vez de precisión. La precisión es lo inverso de la varianza: cuanto más alta es la precisión, más baja es la varianza.

*

Véase <**www.youtube.com/watch?v=vJG698U2Mvo**>.

*

Pensemos, por ejemplo, en el caso de la ascidia. En su estadio joven, este animal simple posee un cerebro bien definido, aunque rudimentario, que usa mientras busca una roca o una acumulación de coral atractiva sobre la que pasarse el resto de la vida alimentándose de lo que extraiga de filtrar el agua que pase por allí. Cuando encuentra ese lugar y se adhiere a él, digiere su propio cerebro y no conserva más que un sistema nervioso simple. (Hay quienes han usado la ascidia como analogía de la carrera profesional académica por la diferencia que parece haber entre el momento anterior y el posterior a conseguir una plaza permanente en una universidad).

*

Al ojo izquierdo se le muestra un patrón cambiante de rectángulos que va perdiendo contraste con el tiempo, mientras que al derecho se le enseña una figura (de una casa o de una cara) que aumenta de contraste progresivamente. Para dirigir cada imagen desde el monitor de un ordenador hasta el ojo correspondiente se utiliza un estereoscopio de espejos. A los participantes se les induce a esperar una cara o una casa enseñando la palabra «cara» o «casa» a ambos ojos al inicio de cada ensayo.

*

La liverpuliana Cilla Black fue una estrella pop en los años sesenta y, posteriormente, una celebridad televisiva en el Reino Unido.

*

Véase una versión en color (y más efectiva) de esta ilusión en <**www.illusionsindex.org/i/rotating-snakes**>.

*

Véase <**www.youtube.com/watch?v=hhXZng6o6Dk**>.

*

A fin de cuentas, aun sin intervención de teletransporte alguno, las células de nuestro cuerpo se están renovando continuamente; la mayoría se han sustituido después de unos diez años, aproximadamente: son una especie de versión biológica del barco de la Paradoja de Teseo. Y no parece que eso tenga especial impacto en nuestro sentido de la identidad personal.

*

Fiódor Dostoievski popularizó las alucinaciones de *doppelgänger* con su novela breve de 1846 *El doble*. Se sabe que Dostoievski padecía epilepsia severa.

*

El hipocampo es una pequeña estructura curvada, ubicada en una zona profunda del lóbulo temporal medial, que se asocia desde hace tiempo con la consolidación de los recuerdos. El nombre procede de la palabra griega con la que se designa el «caballito de mar».

*

Descartes tenía un perro de compañía llamado Monsieur Grat («Señor Rasca») por el que, al parecer, sentía la mayor de las devociones. Aunque, bueno, también viviseccionaba conejos.

*

Entre la exterocepción y la interocepción, está también la propiocepción, que hace referencia a la percepción de la posición y el movimiento del cuerpo (véase el capítulo 5). Es importante no confundir la interocepción con la introspección, que se refiere al examen interior que una persona hace de sus propios estados mentales.

*

El córtex insular o ínsula tiene ese nombre por su parecido con una «isla» situada en medio del ancho «mar» cortical.

*

En septiembre de 2020, hice una excursión de montaña por la temida cresta Sharp Edge de la montaña Blencathra, en el Distrito de los Lagos de Inglaterra. Aunque no se necesita equipo de escalada para esa travesía, cruzar Sharp Edge nunca es fácil. Esa cresta es una agreste mezcla de roca resbaladiza flanqueada por precipicios, y ocurren accidentes. En esta travesía en particular, me fijé en que había una piedra volteada en la base de la cresta en la que alguien había escrito con tiza: «¿Quieres casarte conmigo, Maria?». No pude evitar preguntarme si quien lo hizo estaba al tanto del experimento de Dutton y Aron y estaba intentando aprovecharse de ello.

*

Con este salto directo desde la teoría valorativa hasta la inferencia interoceptiva, estoy omitiendo un amplio corpus de investigaciones intermedias. Antonio Damasio, en particular, ha realizado contribuciones fundamentales que muestran cuál es la relación entre la emoción y la cognición, y cómo ambas dependen del cuerpo. Y Lisa Feldman Barrett llegó por su propia vía independiente a conclusiones muy relacionadas con estas, en las que recalca la importancia de las predicciones interoceptivas.

*

Me estoy refiriendo sobre todo a enfoques surgidos a raíz de la Ilustración. En otros sistemas de creencias anteriores, como el animismo, sí se atribuía intención y finalidad (y vida, y espíritu) a muchas cosas con mucha mayor facilidad.

*

Cabría preguntarse si existe alguna diferencia entre «ser» un modelo y «tener» un modelo. Creo que se puede decir que los sistemas que poseen modelos generativos explícitos capaces de producir predicciones condicionales o contrafactuales, como el Sistema B, «tienen» modelos. Los reguladores que son relativamente fijos e inflexibles, como un termostato por retroalimentación simple como es el Sistema A, tal vez no pasen de «ser» un modelo.

*

De hecho, si te tomas este consejo al pie de la letra, la bola terminará impactándote justo entre los ojos.

*

Se entiende así, también, que habrá una «desatención» sistemática a las señales sensoriales interoceptivas cuando se trate de posibilitar que las «interoacciones» regulen las variables esenciales, exactamente igual que ocurre cuando se atenúan las señales sensoriales propioceptivas durante las acciones externas (como se comentó en el capítulo 5).

*

Hay algunos fascinantes casos límite de entes que no solemos considerar vivos, pero que, aun así, parecen mantener activamente sus identidades: pensemos en el caso de un tornado o de un remolino.

*

¿Cómo puede un estado ser estadísticamente esperado e improbable a la vez? Esto es posible cuando un sistema habita solamente un repertorio restringido o subconjunto de estados –un «conjunto atractor»– respecto a un gran número total de estados posibles. El conjunto atractor es esperado desde el punto de vista estadístico porque es ahí donde se suele encontrar el sistema, pero también es improbable, porque existen muchos estados más fuera del conjunto que dentro de él. Existen muchas más maneras de ser puré que de estar vivo, por así decirlo.

*

En termodinámica, la energía libre es la cantidad de energía disponible para realizar trabajo a una temperatura constante. Es «libre» en el sentido de que está «disponible». La clase de energía libre a la que hace referencia el PEL recibe el nombre de «energía libre variacional», que es un término que proviene del campo del aprendizaje automático y de la teoría de la información, pero que guarda estrecha relación con su equivalente termodinámico.

*

Otro ejemplo de principio parecido al PEL es el «principio de acción estacionaria» de Hamilton, en física, que se puede usar para derivar fórmulas (contrastables) de movimiento e incluso la relatividad general.

*

Los conceptos y las matemáticas que subyacen al PEL no son simples, ni siquiera para expertos en la materia. Así se advierte, por ejemplo, en las primeras líneas de un manual sobre mecánica estadística: «Ludwig Boltzmann, que dedicó gran parte de su vida a estudiar mecánica estadística, se quitó la vida en 1906. Paul Ehrenfest, continuador de su obra, falleció en similares circunstancias en 1933. Y ahora nos toca a nosotros estudiar la mecánica estadística».[10]

*

Se le pide a la persona voluntaria que doble la muñeca cuando ella elija hacerlo y que, al mismo tiempo, señale la posición del punto rotatorio del osciloscopio en el instante preciso en que sienta la intención consciente de mover la mano. Otros dispositivos miden su actividad muscular (EMG) y cerebral (EEG). El panel inferior muestra el promedio típico de los EEG sincronizados con el inicio del movimiento (segundo 0). Las flechas muestran el momento del impulso consciente (A) y el inicio del potencial de disposición (B).

*

A veces, las acciones voluntarias se experimentan también como si requirieran de un esfuerzo consciente, una «fuerza de voluntad». Cuando redacto esta nota a pie de página, por ejemplo, siento que estoy haciendo un esfuerzo. Pero muchas acciones voluntarias iniciadas por el yo precisan de mínimo (o nulo) esfuerzo consciente. De ahí la importancia de no confundir la fuerza de voluntad con (la experiencia del) libre albedrío.

*

Heráclito: «Ningún hombre entra dos veces en el mismo río, porque el río ya no es el mismo y él tampoco».

*

Los sistemas jurídico-legales occidentales se basan en el principio de que la responsabilidad penal requiere tanto de un «acto culpable» *(actus reus)* como de una «mente culpable» *(mens rea)*. Cuando la facultad de una persona para ejercer su libre albedrío –es decir, para controlar sus grados de libertad– se daña o se suprime de algún modo, ¿se puede decir que tiene una «mente culpable»? Algunos –como el filósofo Bruce Waller– sostienen que, puesto que no decidimos tener los cerebros que tenemos, el propio concepto en sí de responsabilidad moral es incoherente.[25] Desde otra perspectiva (que a mí me atrae más), se defiende que, desde el momento en que traspasamos un cierto umbral de competencia para controlar nuestros grados de libertad, se nos *puede* responsabilizar de nuestros actos.

*

Usaré aquí la palabra «animal» para referirme a los animales «no humanos», aunque los seres humanos somos animales también.

*

En un estudio de 2014, se hizo que unas ratas decidieran entre diferentes opciones asociadas a diferentes niveles de recompensa. Cuando escogían una opción que les reportaba una recompensa menor de la que esperaban, aumentaba la probabilidad de que echaran vistazos atrás, hacia la opción no elegida. Los investigadores interpretaron esto como una señal conductual de arrepentimiento, aunque no está ni mucho menos claro qué era lo que de verdad sentían aquellas ratas (si es que sentían algo).

*

Del mundo tal como lo experimenta un animal se suele decir que es el *Umwelt* de ese animal, un término introducido por el etólogo Jakob von Uexküll.

*

Al poco de nuestra visita, Cayo Santiago –junto con buena parte de Puerto Rico– fue devastada por el huracán María. Por fortuna, la mayoría de los monos sobrevivieron. Sin embargo, muchas de las instalaciones de investigación quedaron destruidas. El material que grabamos figura en el documental de 2018 *The Most Unknown* (www.themostunknown.com).

*

Entre los cefalópodos se encuentran los pulpos, los calamares y las sepias, así como otras criaturas relativamente simples, como los nautilos; en total, existen aproximadamente unas ochocientas especies conocidas de ellos. El término «cefalópodo» puede traducirse literalmente del griego como «cabeza-pie», lo que no describe demasiado bien a los pulpos, que, más que pies anexos a la cabeza, tienen apéndices a modo de brazos (tentáculos).

*

Los experimentos no fueron tan terribles como parece. Los pulpos no aparentan notar mucho que se les seccione un tentáculo y, además, estos vuelven a crecer con bastante rapidez. Obviamente, eso no significa que se deban realizar experimentos así sin más, es decir, a menos que haya una muy buena razón para ello.

*

Estas caras sintéticas en particular se generaron con <**thispersondoesnotexist.com**>.

*

Algunas de esas tecnologías no son tan novedosas como parecen. Mi colega Ryota Kanai recordaba hace poco que «un caballo es, básicamente, un caballo de conducción autónoma».

ÍNDICE

PREFACIO
I. NIVEL
1. El problema real
2. Medir la conciencia
3. Fi
II. CONTENIDO
4. Percibir de dentro hacia fuera
5. La magia de las probabilidades
6. La participación del espectador
III. EL YO
7. Delirio
8. Espérate a ti mismo
9. Ser un animal-máquina
10. Pez en el agua
11. Grados de libertad
IV. LOS OTROS
12. Más allá de lo humano
13. Mentes-máquina
EPÍLOGO
AGRADECIMIENTOS
NOTAS
REFERENCIAS BIBLIOGRÁFICAS